Validation de criteres ergonomiques pour l'evaluation d'interfaces utilisateurs

Christian Bastien

UNITE DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél (1) 39 63 55 11

Rapports de Recherche

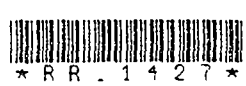
N° 1427

Programme 3
Intelligence artificielle, Systèmes cognitifs et Interaction homme-machine

VALIDATION DE CRITERES ERGONOMIQUES POUR L'EVALUATION D'INTERFACES UTILISATEURS

Christian BASTIEN

Mai 1991



Programme 3

Intelligence Artificielle, Systèmes Cognitifs et Interaction Homme-Machine

(Projet de Psychologie Ergonomique pour l'Informatique)

VALIDATION DE CRITÈRES ERGONOMIQUES POUR L'ÉVALUATION D'INTERFACES UTILISATEURS

A validation of ergonomic criteria for the evaluation of user interfaces

Christian Bastien

Rapport de recherche INRIA N°

VALIDATION DE CRITÈRES ERGONOMIQUES POUR L'ÉVALUATION D'INTERFACES UTILISATEURS

RÉSUMÉ

Cette étude s'inscrit dans un contexte plus général qui est celui de l'organisation des connaissances ergonomiques dans le domaine des interfaces utilisateurs. Cette organisation cherche à rendre ces connaissances plus facilement accessibles et utilisables par les spécialistes et non-spécialistes de l'ergonomie. L'objectif de la présente étude est de valider les définitions de critères ergonomiques pour l'évaluation des interfaces utilisateurs établies lors d'une étude précédente. Pour ce faire, vingt-quatre participant(e)s (12 spécialistes et 12 non-spécialistes de l'ergonomie) prennent part à une tâche d'identification de concept (critères ergonomiques) à partir d'exemples. Cette tâche consiste plus précisément à identifier, pour un problème de conception d'une interface donnée, le critère ergonomique correspondant. La plupart des problèmes de conception sont tirés d'une application réelle permettant la gestion d'une base de données bibliographiques. Les résultats montrent que les deux groupes de participant(e)s ne se distinguent ni par le nombre d'identifications correctes ni par le temps consacré à la lecture des définitions ou à la tâche d'identification. Les identifications correctes vont en moyenne de 59,85 à 74,3% selon qu'il s'agit des critères élémentaires ou principaux. De plus, l'examen détaillé des données permet de déterminer des classes de critères bien définis et des classes de critères dont les définitions bénéficieraient d'améliorations. Chacune de ces classes comporte la moitié des critères élémentaires. En dépit de certains problèmes méthodologiques, cette étude permet d'envisager une méthode d'évaluation basée sur des critères ergonomiques définis de façon explicite et non ambigüe.

Mots-clés : Critères ergonomiques, Évaluation, Interfaces utilisateurs.

A VALIDATION OF ERGONOMIC CRITERIA FOR THE EVALUATION OF USER INTERFACES

ABSTRACT

The aim of the present experiment was to validate a set of ergonomic criteria for the evaluation of user interfaces. This study was performed in the general context of identifying methods and tools for organizing human factors knowledge. This organization aims at making that knowledge more directly and more easily usable by human factors specialists as well as by non-specialists. Criteria definitions that were designed in a previous study were tested in a concept (criteria) identification task. Twenty-four subjects (12 human factors specialists and 12 non-specialists) were asked to identify the appropriate criteria within a set of usability problems. Most of the problems were collected from the interface of an existing bibliographic data base application. The results show no differences between groups of subjects in terms of correct identification. The mean percentages varied from 59.85 to 74.3, for the elementary and the main criteria respectively. The performance times did not vary significantly between groups. In addition, a more detailed examination of the data permits the identification of categories of well defined criteria and categories of criteria that would benefit from an improvement in their definition. Each category contains half of the elementary criteria. Despite some methodological problems, this experiment seems to support the feasibility of an evaluation method based on explicitely defined criteria.

Keywords : Ergonomic criteria, Evaluation, User interface.

TABLE DES MATIÈRES

Page

RÉSUMÉ iii
ABSTRACT iii
TABLE DES MATIÈRES v
LISTE DES FIGURES vii
LISTE DES TABLEAUX vii
1. INTRODUCTION 1
Les méthodes d'évaluation faisant appel aux utilisateurs 1
Les méthodes d'évaluation qui s'appuient sur des modèles théoriques et/ou formels 2
L'évaluation experte 4
Position du problème 5
Objectif de l'étude 7
2. MÉTHODOLOGIE 8
2.1. Participant(e)s 8
2.2. Matériel 8
2.3. Déroulement de l'expérience 10
2.4. Recueil et traitement des données 11
2.4.1. Le temps 11
2.4.2. Scores globaux 11
2.4.3. Qualité des définitions 12
2.4.4. Confusions systématiques entre critères 12
3. RÉSULTATS 13
3.1. Le temps 13
3.2. Scores globaux 13
3.3. Qualité des définitions 15
3.3.1. Critères élémentaires bien définis 16
3.3.2. Critères élémentaires définis de façon satisfaisante 16
3.3.3. Critères élémentaires nécessitant une amélioration de leur définition 17
3.4. Confusions systématiques entre critères 17
3.4.1. Confusions systématiques avec un seul autre critère élémentaire 19
3.4.2. Confusions systématiques avec plusieurs autres critères élémentaires 19
3.5. Exemples d'analyse des confusions systématiques pour modifications ultérieures des définitions 21
4. DISCUSSION 24
RÉFÉRENCES 31
ANNEXES 35
Annexe 1. Définitions, justification et exemples associés aux critères 37
Annexe 2. Matrices de confusions 49

LISTE DES FIGURES

Page

Figure 1. Page écran numéro 1 apparaissant sur la planche 7. 10

Figure 2. Copie d'écran illustrant l'énoncé du problème numéro 54 se rapportant au critère élémentaire "Groupement / Distinction par le format" 22

Figure 3. Copie d'écran illustrant l'énoncé du problème numéro 29 se rapportant au critère élémentaire "Groupement / Distinction par le format" 23

LISTE DES TABLEAUX

Page

Tableau 1. Liste des critères 9

Tableau 2. Nombre moyen d'identifications correctes des critères et coefficients d'accord chez les deux groupes de participant(e)s 14

Tableau 3. Classification des critères élémentaires selon leur fréquence moyenne d'identification et leur proportion d'identification correcte chez les expérimenté(e)s et les inexpérimenté(e)s 15

Tableau 4. Classification des critères élémentaires selon que leur définition doit ou non subir des modifications 17

Tableau 5. Confusions systématiques et non systématiques entre critères élémentaires chez les participant(e)s expérimenté(e)s et inexpérimenté(e)s 18

Tableau 6. Fréquences des identifications incorrectes pour chacun des énoncés dans le cas des confusions systématiques et probabilités associées aux tests binômiaux 20

Tableau 7. Matrice de confusion des critères chez les participant(e)s expérimenté(e)s 51

Tableau 8. Matrice de confusion des critères chez les participant(e)s inexpérimenté(e)s 53

1. INTRODUCTION

Les méthodes d'évaluation des interfaces utilisateurs actuellement disponibles sont aussi variées que nombreuses. Toutes présentent des avantages et des inconvénients : aucune d'elles ne peut prétendre à une évaluation exhaustive de l'interface. Plusieurs types de classification de ces méthodes existent (ex.: Christie et Gardiner, 1990 ; Karat, 1988 ; Maguire et Sweeney, 1989 ; Senach, 1990). On peut distinguer trois grandes catégories de méthodes selon qu'elles font appel : aux utilisateurs, à des modèles théoriques et/ou formels ou aux jugements d'experts[1]. C'est plus particulièrement à ce dernier type d'évaluation que nous nous intéressons. Plus précisément, c'est sur le développement d'une méthode, voire d'outils d'aide à l'évaluation basés sur des critères ergonomiques que portent nos efforts. Cette méthode devrait pouvoir être utilisée par des spécialistes mais aussi par des non-spécialistes de l'ergonomie. Le présent rapport fait état d'une étape essentielle de ce projet, à savoir la validation du jeu de critères ergonomiques proposé par Scapin (1990a, 1990b) qui devrait servir de base à cette méthode.

Avant d'évoquer la question de l'évaluation expérimentale de ces critères, les différentes méthodes d'évaluation et les critiques qu'on leur adresse seront brièvement présentées. L'accent sera mis sur l'évaluation experte, sur les problèmes qu'elle soulève, sur les aides dont disposent actuellement les experts et sur les difficultés d'utilisation de ces dernières.

Les méthodes d'évaluation faisant appel aux utilisateurs

Dans cette catégorie, des données objectives et subjectives peuvent être recueillies. Les premières se rapportent généralement aux performances et aux comportements des utilisateurs lors d'interactions avec un logiciel ou se réfèrent à des indices cognitifs (compréhension et connaissance de l'interface). Les données subjectives quant à elles, sont relatives aux attitudes et/ou aux opinions des utilisateurs. Le recueil ou l'enregistrement de ces données s'effectue à l'aide d'outils dont les effets d'intrusion sont plus ou moins importants. En effet, les méthodes telles que l'observation *in situ*, l'enregistrement vidéo, l'enregistrement physiologique ou encore les verbalisations concomitantes peuvent influencer les performances et le comportement des utilisateurs. Au contraire, les questionnaires, les verbalisations consécutives produites au terme de l'interaction et les entretiens individuels ou de groupe n'altèrent aucunement l'interaction et la performance. Néanmoins, ces

---

[1] D'autres classifications mettent l'accent sur le moment de l'évaluation dans le processus de conception (ex.: Senach, 1990).

techniques posent d'autres problèmes. Ainsi par exemple, les verbalisations[2] consécutives ne donnent pas les mêmes résultats selon qu'elles sont produites avec ou sans trace de l'exécution (ex.: Hoc et Leplat, 1983). Par ailleurs une méthode supposée non-intrusive comme l'enregistrement automatique des actions sur les différents dispositifs de commandes (mouchards électroniques) pose le problème de son acceptabilité par les utilisateurs des logiciels que l'on cherche à évaluer. De plus cette méthode soulève des problèmes techniques tels que son "implémentation", l'espace mémoire nécessaire, le type de système d'opération utilisé et la quantité de données à enregistrer.

Mais le problème qui se pose dès le départ avec ces méthodes est celui de la disponibilité des futurs utilisateurs de l'interface, ou à défaut d'utilisateurs potentiels possédant autant que possible les caractéristiques des premiers. De plus, le temps consacré à l'observation de ces utilisateurs est généralement long, comme l'est celui nécessaire au codage ou décodage des données recueillies, à leur analyse et finalement à leur interprétation. Mentionnons aussi que bon nombre de ces méthodes font appel à des appareils coûteux et complexes. Ces derniers nécessitent souvent une mise en place laborieuse et exige un environnement physique (salles d'observation, d'expérimentation, etc.) et social (spécialistes de différents domaines) dont seuls les grands laboratoires et centres de recherche peuvent se munir. Ces appareils se prêtent donc difficilement à une utilisation en situation naturelle.

En résumé, bien qu'elles permettent d'obtenir des données quantitatives et qualitatives précieuses sur les performances et opinions des utilisateurs, ces méthodes sont difficiles à mettre en place. Des environnements complexes sont nécessaires et des précautions particulières doivent être prises afin que les données ne soient pas trop influencées par le caractère intrusif de ces méthodes. Par ailleurs la quantité de données qu'elles produisent est difficile à analyser et à interpréter. Ces raisons rendent l'utilisation de ces méthodes en situation naturelle extrêmement difficile.

Les méthodes d'évaluation qui s'appuient sur des modèles théoriques et/ou formels

Les évaluations qui s'appuient sur des modèles théoriques et/ou formels permettent de prédire la complexité d'un système et par conséquent les performances des utilisateurs. Ces modèles sont particulièrement utiles lorsqu'une évaluation ergonomique ne peut avoir recours à ces derniers. On trouve différents types de modèles : les modèles de tâches (ex.: KLM et GOMS de Card, Moran et Newell,

---

[2] A propos des verbalisations, voir les commentaires de Hayes (1986) et la revue de question de Caverni (1988).

1983) ; les modèles linguistiques (ex.: ALG de Reisner, 1983 ; CLG de Moran, 1981) ; et les modèles cognitifs de l'interaction (ex.: CCT de Kieras et Polson, 1985). Les premiers prédisent les durées d'exécution ou les occurrences d'erreurs à partir de la décomposition des tâches complexes en unités élémentaires et du temps de réalisation de celles-ci. Cependant, soit ces modèles sont simplificateurs, ce qui limite leur intérêt pour prédire les performances en situations naturelles ; soit encore la modélisation des tâches est très longue, ce qui est un handicap sérieux compte tenu des exigences temporelles de développement des logiciels (Green, Schiele et Payne, 1988). Pour leur part, les modèles linguistiques tentent de rendre explicite, sous forme d'une grammaire, la structure de l'interface. Un de ceux-ci (le modèle ALG), construit un modèle des actions mis en jeu par l'utilisateur lors de l'exécution d'une tâche. En d'autres termes, cette grammaire décompose récursivement les buts de l'utilisateur pour aboutir aux actions élémentaires nécessaires à leur atteinte. Les règles de cette grammaire établissent la correspondance entre les buts et les opérations à mettre en œuvre. Dans ce modèle, le nombre d'actions élémentaires différentes pour atteindre un but, la longueur des séquences d'actions pour une tâche donnée, le nombre de règles non nécessaires et le nombre de règles pour les séquences terminales similaires fournissent respectivement des indices sur la complexité du langage, la simplicité des procédures et la cohérence de la structure. Or il semble que les critères avancés pour évaluer la complexité d'un langage de commande ne soient pas suffisants. La facilité d'utilisation ne semble pas être une simple fonction du nombre de règles et de la longueur des règles : des tâches faciles à décrire peuvent être difficiles à réaliser. Par ailleurs Senach (1990) souligne que la description est trop indépendante de l'environnement matériel. Aucune référence n'est faite au mode de dialogue ou encore aux affichages. Un autre modèle (CLG), utilisé surtout pour la conception mais permettant aussi l'évaluation, décompose l'interface en différents niveaux d'abstraction (niveau conceptuel : concepts abstraits et tâche : niveau de la communication : langage de commande ; niveau physique : affichage, entrées de données, etc.). Chacun de ces niveaux se décompose ensuite en sous-niveaux permettant une description complète de l'interface. La principale critique qu'on adresse à cette méthode est qu'elle n'est pas applicable directement à l'évaluation d'une interface développée avec une méthode autre que CLG (Senach, 1990).

Les modèles cognitifs et plus particulièrement le modèle CCT permettent une simulation de l'interaction entre l'utilisateur et le dispositif à partir de trois éléments : un modèle de l'utilisateur exprimé sous forme d'un système de production ; une représentation de la tâche utilisant la notation du modèle GOMS ; et un modèle de l'interface. La complexité de l'interface est ici fonction de la quantité, du contenu et de la structure des connaissances requises pour utiliser efficacement l'interface. Elle est évaluée par des indicateurs tels que le nombre total de règles de production nécessaires

à la modélisation de la tâche, le nombre de productions déclenchées, etc. Cependant tous les auteurs ne s'entendent pas sur l'interprétation de la complexité et sur la façon de l'évaluer. Pour certains, la complexité se reflèterait davantage par la mise en jeu de sources différentes de connaissance que par le nombre de règles de production. Par ailleurs, l'application de ce modèle serait limitée aux tâches n'impliquant pas de résolution de problèmes. Finalement ce modèle ne serait pas encore validé empiriquement (Knowles, 1988).

En résumé, la plupart des techniques précédentes, bien qu'elles constituent des outils intéressants d'évaluation, semblent insuffisantes pour constituer des méthodes d'évaluation à part entière. Elles ne semblent pas pouvoir fournir une évaluation complète de l'interface, de plus elles apparaissent comme étant très coûteuses. Elles nécessitent beaucoup de temps et apparaissent trop difficiles à mettre en œuvre et ce, surtout par les non-spécialistes (Bellotti, 1988). Enfin, elles sont surtout comparatives : elles permettent d'évaluer différentes versions d'une même interface les unes par rapport aux autres mais pas de juger de la qualité globale d'une interface. Ce sont probablement les raisons qui expliquent pourquoi bon nombre d'évaluations reposent sur l'expertise.

L'évaluation experte

L'évaluation à laquelle procède l'expert peut consister à comparer les performances d'un système aux recommandations ou standards existants, ou encore elle peut consister en une évaluation plus subjective (Maguire et Sweeney, 1989). Cette dernière repose alors sur des formations académiques, sur des expériences accumulées sur le terrain et sur des examens de nombreuses données expérimentales qui peuvent être fort différentes. Ces évaluations sont avantageuses si on les compare aux méthodes précédentes car elles sont relativement peu coûteuses, elles se font assez rapidement et peuvent survenir assez tôt dans le processus de conception.

Cependant, les études effectuées sur l'activité des ergonomes en situation d'évaluation montrent que les performances individuelles sont très variables quant au nombre et au type de problèmes évoqués (Hammond, Hinton, Barnard, MacLean, Long et Whitefield, 1985 ; Molich et Nielsen, 1990 ; Nielsen et Molich, 1990 ; Pollier, 1991) et quant aux stratégies d'évaluation adoptées (Pollier, 1991). Toujours d'après ces études, les évaluateurs auraient tendance à se focaliser sur des aspects particuliers des interfaces et la communauté des problèmes détectés serait relativement faible. Seule une synthèse des différentes évaluations permettrait d'obtenir une évaluation complète de l'interface. La détection des erreurs, même pour des interfaces simples, serait donc une activité relativement difficile (Molich et Nielsen, 1990).

Position du problème

Les outils ou aides dont dispose actuellement l'ergonome ou le concepteur pour l'évaluation des interfaces apparaissent sous forme de standards, de guides, de règles, d'algorithmes (voir à ce propos Smith, 1988) et de listes de contrôle (*checklists*).

Le meilleur exemple de guide et sans doute le plus cité, est celui de Smith et Mosier (1986) ou ses versions précédentes. Ce document, source de référence pour la plupart des guides postérieurs à sa publication (e.g. : Brown, 1988 ; Ravden et Johnson, 1989 ; Scapin, 1986 ; Shneiderman, 1987 ; etc.), est une compilation de plusieurs centaines de recommandations (944 au total). Les autres guides répondent à des objectifs divers et ont en commun de présenter un nombre de recommandations ou de règles beaucoup plus restreint que celui observé dans celui de Smith et Mosier (1986). Certains abordent plus spécifiquement la conception des interfaces utilisateurs bien qu'ils puissent aussi être utilisés pour l'évaluation (e.g. : Brown, 1988 ; Gardiner et Christie, 1987 ; Heckel, 1984 ; Rivlin, Lewis et Cooper, 1990 ; Rubinstein et Hersh, 1984 ; Scapin, 1986 ; Shneiderman, 1987) alors que d'autres traitent principalement de l'évaluation. Ceux-ci peuvent être plus ou moins complexes et plus ou moins détaillés. On trouve des chapitres de livre, succincts et assez généraux (ex.: Christie et Gardiner, 1990 ; Gardner, Mayfield et Maguire, 1985 ; Marshall, Nelson et Gardiner, 1987 ; Polson, 1988) ou encore des guides plus détaillés comportant des listes de contrôle (e.g. : Clegg, Warr, Green, Monk, Kemp, Allison et Landsdale, 1988 ; Oppermann, Murchner, Paetau, Pieper, Simm et Stellmacher, 1989 ; Ravden et Johnson, 1989).

Il existe très peu de données sur l'utilisation et l'utilité de ces guides. Les seules données existantes sont celles qu'ont recueillies Mosier et Smith (1985, 1986). Ces auteurs ont montré que leur guide s'avère utile. Les personnes à qui ils l'ont envoyé l'ont lu, l'ont utilisé, envisagent de l'utiliser de nouveau et de le recommander à d'autres personnes. Les auteurs constatent que les recommandations qu'il contient sont surtout utilisées au début de la conception bien qu'elles soient aussi utiles pour l'évaluation des systèmes existants ou en cours de conception. Toutefois, les recommandations, si elles sont faciles à utiliser par les ergonomes, le sont beaucoup moins par les concepteurs. Un certain degré d'expertise du domaine serait nécessaire pour être en mesure de bien les interpréter (Scapin, 1990b). Une autre difficulté est liée au fait que les recommandations sont accompagnées de très peu d'informations sur la façon de les utiliser (Ravden et Johnson, 1989). Leur recherche devient vite un casse-tête et rien ne permet de trancher lorsque les recommandations trouvées sont contradictoires.

Ces difficultés d'utilisation de la compilation de Smith et Mosier (1986) ne sont probablement pas sans liens avec le développement parallèle d'autres guides. Ce qui frappe, lorsqu'on regarde ceux-ci, est l'absence d'uniformité dans la présentation du contenu. Les recommandations qu'on y trouve sont présentées soit sous divers critères, principes ou thèmes de plus hauts niveaux qui tentent de les organiser et de les synthétiser (e.g. : homogénéité, compatibilité stimulus-réponse, facilité d'apprentissage), soit sous des thèmes issus d'un découpage de l'interface (e.g. : langage de commande, sélection de menus) les principes et critères faisant l'objet d'une présentation indépendante. L'hétérogénéité de ces guides semble attester de la difficulté que pose l'organisation des connaissances sans cesse croissantes[3] du domaine de l'interaction personne-ordinateur (IPO).

La plupart de ces guides présentent un certain nombre de critères ergonomiques. Ce nombre n'est cependant pas le même d'un auteur à l'autre et un critère donné n'est pas nécessairement défini de la même façon par tous les auteurs qui l'utilisent. Il arrive que des critères différents se réfèrent au même contenu et inversement que des critères identiques se réfèrent à des contenus différents. Néanmoins, les critères semblent constituer une solution intéressante au problème de l'organisation des connaissances et plus particulièrement des recommandations. De plus, s'ils étaient stables et clairement définis ils pourraient être utilisés pour : la formation à l'ergonomie ; la constitution de grilles d'évaluation et la structuration des rapports d'évaluation ; la définition de standards ; et enfin, la recherche dans des bases de données ergonomiques, l'implémentation de règles et la définition de métriques (Scapin, 1990a, 1990b).

La variabilité du nombre de critères et de la précision de leur définition est probablement fonction du processus d'organisation qui leur a donné lieu. Trois stratégies de conception semblent à l'origine des guides actuels. Une de celles-ci consiste à interroger les connaissances acquises dans les domaines de recherche de la psychologie cognitive tels que le raisonnement, les modèles mentaux, la mémoire, le langage et l'acquisition des habiletés, même si ces connaissances ne sont pas directement liées aux interfaces utilisateurs, pour en extraire des recommandations qui soient applicables au domaine de l'IPO. Des critères ou dimensions-clés sont ensuite utilisés pour désigner les regroupements de ces recommandations (ex.: Marshall, Nelson et Gardiner, 1987). Une deuxième approche consiste à faire une revue des critères utilisés (ex.: Johnson, Clegg et Ravden, 1989 ; Ravden, 1988 ; Ravden et Johnson, 1989). Cette démarche s'appuie donc sur des synthèses ou revues de question

---

[3] Il n'y a, pour se rendre compte de cette croissance, qu'à consulter l'Ergonomics Abstract publié mensuellement par le "Ergonomics Information Analysis Center" (EIAC).

déjà publiées. La troisième approche est plus empirique en ce sens qu'elle procède des données et des conclusions de recherches à partir desquelles on peut énoncer des recommandations. Ces dernières sont ensuite traduites sous forme de règles puis regroupées et étiquettées à l'aide de critères (ex.: Scapin, 1990b). Cette démarche peut aussi s'enrichir de considérations basées sur des pratiques courantes.

Cette dernière méthode vise l'organisation des connaissances sans cesse croissante pour les rendre facilement utilisables par les non-spécialistes autant que par les spécialistes. Il ne s'agit donc pas seulement de présenter un certain nombre de critères mais bien d'utiliser ceux-ci comme dimensions permettant de rendre compte des recommandations. A cette fin Scapin (1990b) a procédé à un examen des diverses recommandations de la littérature. Cet examen s'est armé d'un mécanisme de décryptage permettant, par le moyen de schémas génériques, de traduire ces recommandations sous forme de règles de production. Un premier jeu de critères, a été proposé afin de classifier ces règles. Une certaine difficulté d'affectation des critères est apparue lors de cette première phase. Les critères ont été affinés en sous-critères ce qui a produit un second jeu de critères. Ce deuxième jeu de critères a permis de prendre en compte un plus grand nombre de recommandations et de classifier les règles de façon plus univoque.

Objectif de l'étude

L'étude de Pollier (1991) a déjà montré que le jeu de critères proposé par Scapin (1990a, 1990b) pouvait servir de grille de description des performances d'évaluation d'interfaces. Toutefois, si ces critères doivent être utilisés par des non-spécialistes, la première tâche consiste à s'assurer qu'ils sont clairement définis, suffisamment justifiés et accompagnés d'exemples qui permettent de bien les comprendre. C'est précisément à cette tâche que s'emploie la présente recherche.

Il s'agit donc ici de valider les critères et leur définition. S'ils sont valides, ces derniers devraient pouvoir être utilisés de la même façon par des personnes ayant des formations et des expériences variées. On peut penser par exemple qu'ils permettraient des résultats semblables à des tâches de classification d'erreurs ou de problèmes de conception, qu'ils permettraient d'uniformiser la présentation des rapports d'évaluation et finalement qu'ils pourraient permettre des évaluations d'interfaces plus cohérentes.

Les critères proposés constituent des concepts. Comme pour tout concept, ceux-ci sont définis, expliqués et accompagnés d'exemples. On dispose de plusieurs méthodes pour évaluer la compréhension des concepts enseignés. On peut par exemple demander une définition du concept, l'identification d'exemples s'y rapportant, la production

d'exemples non enseignés, l'explication du concept, etc. (Gropper, 1983). Le but n'étant pas ici d'évaluer la mémorisation des concepts mais bien la compréhension de ceux-ci, les critères et leur définition seront à la portée des participant(e)s en tout temps. La tâche qui est utilisée dans cette recherche est une tâche d'identification ou de reconnaissance de concept à partir d'exemples s'y rapportant. Si les concepts sont suffisamment explicites et clairement définis, les participant(e)s devraient être en mesure, face à des exemples, de les identifier correctement. Afin de déterminer si cette validité est indépendante de la formation et/ou de l'expérience, deux groupes de participant(e)s se distinguant sur cet aspect seront étudiés.

2. MÉTHODOLOGIE

2.1. Participant(e)s

Vingt-quatre personnes participent à la recherche, dont 12 participant(e)s expérimenté(e)s (ergonomes) et 12 inexpérimenté(e)s (étudiant(e)s). Les premiers, ont en moyenne 7,5 années d'expérience (*E.T.* = 6,5 ; *Mode* = 3) en évaluation et/ou conception d'interfaces utilisateurs. Tous/toutes ont une formation supérieure allant du diplôme de maîtrise à la thèse de psychologie. Ils/elles exercent leur profession dans des sociétés de services, des entreprises ou dans des universités. Les étudiant(e)s sont inscrits au Diplôme d'Etudes Supérieures Spécialisées d'ergonomie (D.E.S.S.) et n'ont aucune expérience en évaluation et/ou conception d'interfaces utilisateurs. Chaque groupe est constitué d'un nombre égal d'hommes et de femmes.

2.2. Matériel

Le matériel expérimental comprend une liste de critères ergonomiques pour la conception et l'évaluation d'interfaces utilisateurs, des documents présentant les définitions et un ensemble d'exemples de problèmes de conception.

La liste de critères ergonomiques (Scapin, 1990a, 1990b) comporte trois niveaux de critères. Le premier niveau est constitué de huit *critères principaux*. Cinq de ceux-ci se subdivisent en sous-critères (critères de deuxième niveau) dont certains se subdivisent à leur tour en sous-critères (critères de troisième niveau). Certains *critères principaux* ne comportent donc aucun sous-critères alors que d'autres peuvent se décomposer en deux autres niveaux de critères. Pour simplifier l'utilisation de la liste des critères, contourner les références aux divers niveaux et faciliter le traitement des données, on utilise le terme *critère élémentaire* pour désigner les critères ou sous-critères qui ne se décomposent pas. Le tableau 1 indique ces derniers. Au total, la liste en contient 18.

Tableau 1

Liste des critères (*)

1. Guidage
    - 1.1. **Prompting**
    - 1.2. Groupement/Distinction entre items
        - 1.2.1. **Groupement/Distinction par la localisation**
        - 1.2.2. **Groupement/Distinction par le format**
    - 1.3. **Feed-back immédiat**
    - 1.4. **Clarté**
2. Charge de travail
    - 2.1. **Brièveté**
        - 2.1.1. **Concision**
        - 2.1.2. **Actions minimales**
    - 2.2. **Charge mentale**
3. Contrôle explicite
    - 3.1. **Actions explicites**
    - 3.2. **Contrôle utilisateur**
4. Adaptabilité
    - 4.1. **Flexibilité**
    - 4.2. **Prise en compte de l'expérience de l'utilisateur**
5. Gestion des erreurs
    - 5.1. **Protection contre les erreurs**
    - 5.2. **Qualité des messages**
    - 5.3. **Correction des erreurs**
6. **Homogénéité / Consistence**
7. **Signifiance des codes**
8. **Compatibilité**

* Les critères élémentaires apparaissent en caractères gras

Deux documents accompagnent cette liste. Le premier donne une définition, des justifications et des exemples pour chaque critère (voir annexe 1). Le deuxième document ne comporte que les définitions, les justifications et exemples ayant été enlevés.

Les exemples de problèmes de conception d'une interface utilisateur sont au nombre de trente-six, soit deux problèmes par *critère élémentaire*. Certains de ces exemples ont été choisis parmi une liste de problèmes observés dans une application permettant la gestion d'une base de données bibliographiques disponible sur Macintosh. Le choix s'est fait sur la base de leur adéquation vis-à-vis des *critères élémentaires*, adéquation évaluée par trois expérimentateurs. Tous les critères n'étant pas transgressés par ce logiciel, quelques exemples d'erreur ont du être créés.

Les problèmes sont énoncés de façon à ce qu'aucun des termes utilisés ne se rapporte directement aux critères. Chaque problème est présenté sur une fiche et est illustré par des copies de pages écran tirées du logiciel bibliographique. Ces pages écran ont été utilisées telles quelles lorsqu'elles illustraient directement les problèmes choisis, ou ont été modifiées pour correspondre aux problèmes créés. Les copies sont

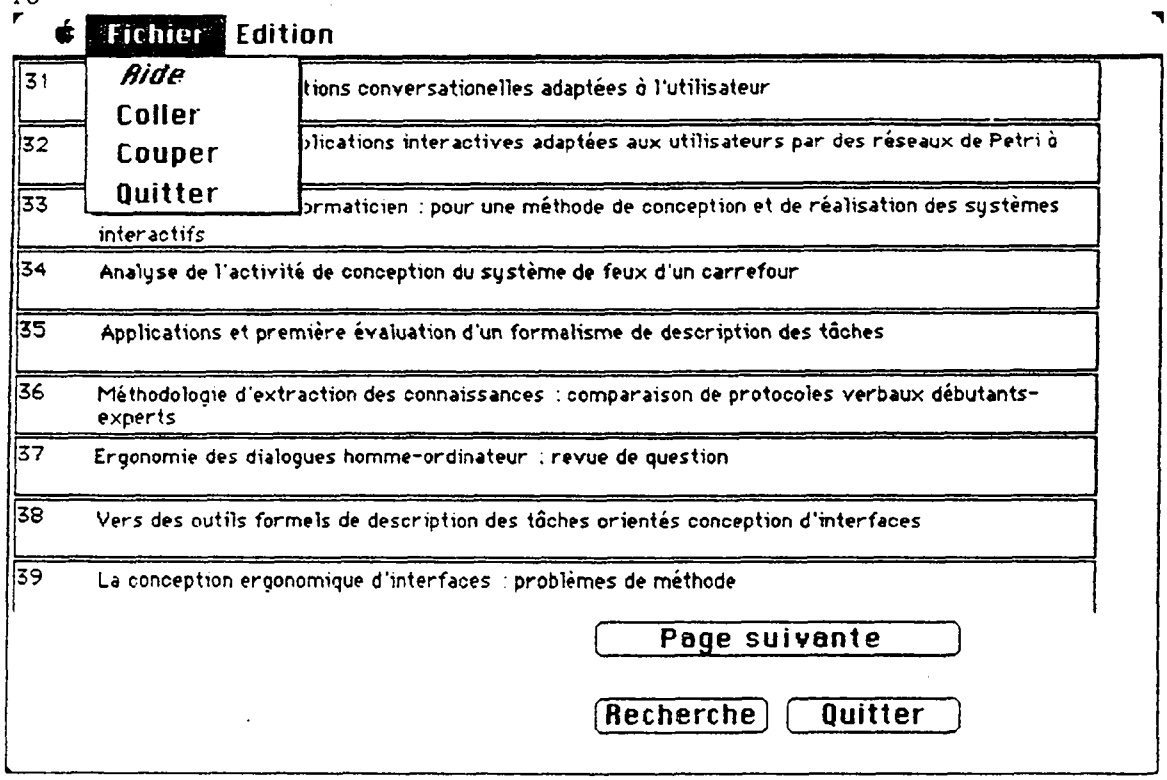


Figure 1. Page écran numéro 1 apparaissant sur la planche 7.

numérotées et présentées aux participants dans des chemises cartonnées (planches). De plus, l'énoncé des problèmes spécifie clairement la ou les pages écrans auxquelles il se rapporte. L'énoncé suivant est un exemple de problème relatif au critère "*Groupement / Distinction par la localisation*" et est illustré par la copie d'écran de la figure 1 : "Dans le menu "**Fichier**" de l'écran **1** apparaissent des options d'édition ("**Couper**" et "**Coller**")".

2.3. Déroulement de l'expérience

Tous/toutes les participant(e)s prennent part, individuellement, à une seule séance expérimentale. En début de séance la consigne écrite leur est donnée. Elle explique le but de la recherche et la tâche à effectuer.

L'expérience se déroule en deux phases. Au cours de la première, les participant(e)s sont invité(e)s à prendre connaissance de la liste des critères et du document présentant les définitions, les justifications et les exemples. L'expérimentateur demande aux participant(e)s de porter une attention particulière aux justifications et exemples et les informe qu'ils/elles n'auront accès, dans la deuxième phase de la recherche, qu'aux seules définitions. Cette première phase terminée, on procède à la phase d'identification des critères. L'expérimentateur présente un à un les problèmes et les chemises contenant les pages écran correspondantes aux

participant(e)s. Ces dernier(e)s doivent choisir, dans la liste, le *critère élémentaire* auquel se rapporte l'énoncé. En d'autres termes, les participant(e)s doivent indiquer le critère ergonomique élémentaire qui n'a pas été pris en compte lors de la conception ; cette lacune ayant entraîné l'erreur présentée. Les participant(e)s ne doivent indiquer qu'un seul *critère élémentaire*. Si en dépit de la consigne certain(e)s participant(e)s identifient plus d'un critère, l'expérimentateur leur demande d'indiquer celui qui leur semble le plus important. Dans l'exemple précédent, *le critère théorique*, i.e. celui donné par les expérimentateurs, est le critère "*Groupement / Distinction par la localisation*" puisque les options "Couper" et "Coller" sont plus appropriées dans le menu "Edition". Une fois le critère identifié, l'expérimentateur reprend la fiche de l'énoncé et les copies d'écran et présente le problème suivant. Cette procédure ne permet pas aux participant(e)s : de revenir sur leur choix ; d'identifier les critères déjà cités par rapport à ceux non cités ; et de découvrir le nombre d'énoncés présentés par critère. Pour chaque participant(e) l'ordre de présentation des erreurs est déterminé à partir de tables de nombres au hasard. Aucune contrainte de temps n'est imposée.

2.4. Recueil et traitement des données

Deux types de données sont recueillies tout au long de l'expérience. Il s'agit du temps d'exécution de la tâche et des réponses des participant(e)s, i.e. les *critères élémentaires* qu'ils/elles identifient à chacun des énoncés.

2.4.1. Le temps

Les temps consacrés à la lecture des critères et de leur définition, puis à la tâche d'identification sont chronométrés. Ces données sont soumises à un test *t* (bilatéral) afin de comparer la performance des participant(e)s expérimenté(e)s et inexpérimenté(e)s.

2.4.2. Scores globaux

Chaque *critère élémentaire* identifié par les participant(e)s est noté. Les réponses données par les participant(e)s sont ensuite comparées aux *critères théoriques*. Cette comparaison permet le calcul du *score global*, i.e. du nombre de *réponses correctes* ou si l'on veut le nombre de réponses identiques aux *critères théoriques*. Un test *t* (bilatéral) permet de comparer les scores des deux groupes de participant(e)s.

Un autre *score global* est calculé, mais cette fois-ci à partir de la comparaison des *critères principaux*, auxquels se réfèrent les *critères élémentaires* identifiés par les participant(e)s, avec les *critères élémentaires* théoriques. Ce second score permet d'apprécier la proportion des erreurs relatives aux difficultés de discrimination entre *critères élémentaires* et celles relatives aux difficultés de discrimination entre *critères*

principaux. Ce score est aussi soumis à un test *t* (bilatéral) afin de comparer les deux groupes de participant(e)s.

Bien que les *scores globaux* donnent une bonne idée de la performance générale des participant(e)s, ils sont peu informatifs sur la nature des erreurs commises et sur les fréquences d'identification correctes et incorrectes des *critères élémentaires*. Pour combler cette lacune, des matrices de confusions sont construites pour chaque groupe (voir annexe 2). Ces matrices permettent le calcul du coefficient d'accord inter-juge *Kappa* (*K*) de Cohen (1960 ; voir aussi Bakeman et Gottman, 1986) et permettent de faire apparaître les *confusions systématiques* entre *critères élémentaires*.

L'indice *Kappa* est calculé à l'aide de la formule suivante :

$$K = (P_O - P_C) / (1 - P_C) \text{ où :}$$

P_O = La proportion des accords observés. Cette proportion s'obtient en additionnant les fréquences se trouvant sur la diagonale de la matrice de confusion (les accords) et en divisant cette somme par le nombre total d'occurrences, soit 432 (36 réponses x 12 participant(e)s).

P_C = La proportion des accords dus au hasard. Elle s'obtient en additionnant les produits croisés des proportions d'accord relatives aux lignes et aux colonnes et en divisant ce nombre par le carré du nombre total d'occurrences soit 186624 pour les matrices concernées.

2.4.3. Qualité des définitions

A partir des fréquences totales d'identification des *critères élémentaires*, une proportion d'identification correcte (p) et une fréquence moyenne d'identification (f) sont calculées. La proportion d'identification correcte peut varier de 0 à 1, ce dernier chiffre représentant les cas où l'identification du critère se fait sans erreur. Cette proportion a été divisée en trois catégories équivalentes : la première allant de 0 à 0,33 ; la deuxième allant de 0,33 à 0,66 ; et la troisième de 0,66 à 1. Quant à la fréquence moyenne d'identification, elle devrait théoriquement être égale à 2, chaque *critère élémentaire* ne devant être identifié que deux fois. Or on doit s'attendre, s'il y a des erreurs d'identification à ce que la moyenne soit inférieure ou supérieure à ce nombre. Dans le premier cas le *critère élémentaire* sera sous-identifié et dans l'autre sur-identifié. *La qualité des définitions* des critères sera évaluée sur la base de ces deux indices : la proportion d'identification correcte (p) et la fréquence moyenne d'identification (f). On distinguera trois catégories de critères : *les critères bien définis* ($p \geq 0,66 \cap f \leq 2$) ; *les critères définis de façon satisfaisante* ($0,33 \leq p < 0,66 \cap f \leq 2$) ; et *les critères nécessitant une amélioration de leur définition* ($0 \leq p < 0,33 \cup f > 2$).

2.4.4. Confusions systématiques entre critères

Les *confusions systématiques* sont définies par des identifications incorrectes d'un même *critère élémentaire* par au moins trois participant(e)s (soit 25% des participant(e)s d'un groupe). Pour déterminer si ces confusions résultent des définitions plutôt que des énoncés de problèmes, une comparaison du nombre total des identifications incorrectes pour chacun des deux énoncés se rapportant à un même *critère élémentaire* est effectuée à l'aide du test binômial. De plus, ces *confusions systématiques* peuvent être uniques ou multiples. Les critères appartenant à l'une ou l'autre catégorie sont identifiés.

Les modifications des définitions s'appuieront sur l'analyse de ces *confusions systématiques* qui permettront d'identifier les *critères élémentaires* difficiles à discriminer. Une attention particulière sera aussi portée aux énoncés de problèmes de façon à s'assurer que les erreurs commises ne résultent pas d'interprétations alternatives possibles de ceux-ci. Deux exemples illustrant ce travail seront présentés.

3. RÉSULTATS

3.1. Le temps

Le temps que consacre les participant(e)s expérimenté(e)s à la lecture des critères et de leurs définitions (M = 14,42 min., $E.T.$ = 4,36) ne diffère pas de façon significative du temps consacré à cette tâche par les participant(e)s inexpérimenté(e)s (M = 17,83 min., $E.T.$ = 7,49) (t (22) = -1,37 , p = 0,19). Aucune différence significative ne s'observe par ailleurs dans le temps accordé à l'identification des critères (t (22) = 0,35 , p = 0,73). Les expérimenté(e)s y passent en moyenne 57,42 min. ($E.T.$ = 14,85) et les inexpérimenté(e)s 55,33 min. ($E.T.$ = 14,76).

3.2. Scores globaux

Les scores globaux montrent que les performances des participant(e)s sont supérieures à 50% et ce, quel que soit le mode de calcul : *critères élémentaires*, *critères principaux*, *coefficients Kappa*. Les identifications correctes peuvent même atteindre près de 78% dans le cas des *critères principaux*. Ces résultats s'avèrent particulièrement intéressants si l'on tient compte du peu de temps que prennent les participant(e)s lors de la phase d'apprentissage. La comparaison des scores obtenus pour les *critères élémentaires* et *principaux* indique par ailleurs que toutes les erreurs d'identifications ne peuvent être réduites aux difficultés discriminatives que posent les définitions des premiers. Les erreurs d'identification résultent aussi de problèmes de discrimination entre *critères principaux*. Le tableau 2 résume les résultats.

Tableau 2

Nombre moyen d'identifications correctes des critères et coefficients d'accord chez les deux groupes de participant(e)s.

		Critères	
Participant(e)s		*Élémentaires*	*Principaux*
Expérimenté(e)s	*Moy*	22,92 (63,7%)	27,92 (77,6%)
	E.T.	4,34	3,34
	K	0,61	
Inexpérimenté(e)s	*Moy*	20,5 (56%)	25,67 (71%)
	E.T.	3,9	4,16
	K	0,51	
	MOY	21,71 (59,85%)	26,79 (74,3%)

Critères élémentaires. Le nombre moyen d'identifications correctes des *critères élémentaires* ne diffère pas de façon significative d'un groupe à l'autre (t (22) = 1,44 , p = 0,17). Les participant(e)s expérimenté(e)s identifient correctement les *critères élémentaires* dans 63,7% des cas (M = 22,92/36 ; $E.T.$ = 4,34). Les participant(e)s inexpérimenté(e)s réussissent quant à eux dans 56% des cas (M = 20,5/36 ; $E.T.$ = 3,9).

Critères principaux. Le pourcentage moyen d'identifications correctes de ces critères est de 77,6% (27,92/36 ; $E.T.$ = 3,34) chez les participant(e)s expérimenté(e)s et de 71% (25,67/36 ; $E.T.$ = 4,16) chez les participant(e)s inexpérimenté(e)s. Ces performances sont significativement supérieures à celles obtenues pour les *critères élémentaires* et ce chez les deux groupes (expérimenté(e)s : t (11) = - 5 , p = 0,0001 ; inexpérimenté(e)s : t (11) = - 5,17 , p = 0,0001). Ces performances ne diffèrent cependant pas d'un groupe à l'autre (t (22) = 1,46 , p = 0,16).

Coefficients Kappa. Les coefficients Kappa indiquent que les expérimenté(e)s (K = 0,61) sont, dans l'ensemble, un peu plus en accord avec les expérimentateurs que ne le sont les inexpérimenté(e)s (K = 0,51). En d'autres termes, le premier groupe confond un peu moins les *critères élémentaires* que le second groupe. Lorsqu'on exprime ces coefficients sous forme de pourcentage plutôt que de proportion, ce qui permet de les comparer aux *scores globaux* relatifs aux *critères élémentaires*, on constate qu'ils sont légèrement inférieurs à ces derniers. Cette situation s'explique par le fait que le calcul des coefficients tient compte de la proportion des accords pouvant résulter du hasard.

Tableau 3

Classification des critères élémentaires selon leur fréquence moyenne d'identification et leur proportion d'identification correcte chez les expérimenté(e)s et les inexpérimenté(e)s. (*)

Proportion d'identifications correctes	Fréquence moyenne d'identification ≤2	Fréquence moyenne d'identification >2
Expérimenté(e)s		
0,66≤p	**Feed-back** **Protection contre erreurs** **Qualité messages** Gr./Dist localisation Clarté Charge mentale Actions explicites Expérience utilisateur Corrections des erreurs Signifiance des codes	**Homogénéité**
0,33≤p<0,66	**Gr./Dist. par le format** Contrôle utilisateur	**Flexibilité** Actions minimales Compatibilité
0≤p<0,33	Concision	Prompting
Inexpérimenté(e)s		
0,66≤p	**Feed-back** **Protection contre erreurs** **Qualité messages** Compatibilité	**Homogénéité** Clarté
0,33≤p<0,66	**Gr./Dist. par le format** Gr./Dist localisation Concision Actions minimales Charge mentale Actions explicites Expérience utilisateur Corrections des erreurs	**Flexibilité** Prompting Contrôle utilisateur Signifiance des codes
0≤p<0,33		

(*) Les critères en caractères gras sont dans la même classe pour les deux groupes.

3.3. Qualité des définitions

Le tableau 3 présente la classification des critères à partir de leur fréquence moyenne d'identification et de leur proportion d'identification correcte selon les groupes. On constatera tout d'abord que la classification des critères dans les six classes n'est pas la même pour les deux groupes. Ce point est important car si les groupes ne

diffèrent pas quant à leurs *scores globaux*, les erreurs commises ne se rapportent pas aux mêmes critères. En d'autres termes, les participant(e)s n'utilisent pas à tort les mêmes *critères élémentaires*. Ainsi chez les expérimenté(e)s par exemple le critère "*Signifiance des codes*" se trouve dans la classe des critères ayant une fréquence moyenne d'identification inférieure ou égale à 2 et une proportion d'identifications correctes plus grande ou égale à 0,66. Chez les inexpérimenté(e)s, ce critère se trouve dans la classe où les fréquences moyennes sont supérieures à 2 et où la proportion des identifications correctes varie entre 0,33 et 0,66.

A partir de ces six classes, trois catégories de critères sont isolées à savoir les *critères élémentaires bien définis* ($p \geq 0,66 \cap f \leq 2$) ; les *critères élémentaires définis de façon satisfaisante* ($0,33 \leq p < 0,66 \cap f \leq 2$) ; et les *critères élémentaires nécessitant une amélioration de leur définition* ($0 \leq p < 0,33 \cup f > 2$).

3.3.1. Critères élémentaires bien définis

Les *critères élémentaires* qui composent cette catégorie sont considérés comme étant bien définis pour l'un ou l'autre groupe ou encore pour les deux. Il s'agit des critères "*Feedback*", "*Protection contre les erreurs*" et "*Qualité des messages*". Les critères "*Gr./Dist. par la localisation*", "*Clarté*", "*Charge mentale*", "*Actions explicites*", "*Prise en compte de l'expérience de l'utilisateur*", "*Correction des erreurs*" et "*Signifiance des codes*" ne semblent pas poser de problèmes pour les expérimenté(e)s. Le critère "*Compatibilité*" semble mieux compris par les inexpérimenté(e)s que par les expérimenté(e)s.

3.3.2. Critères élémentaires définis de façon satisfaisante

Dans cette catégorie on trouve un seul *critère élémentaire* commun aux deux groupes. Il s'agit du critère "*Gr./Dist. par le format*". Par ailleurs, on trouve un seul autre critère chez les expérimenté(e)s et sept autres chez les inexpérimenté(e)s. Cette proportion est exactement l'inverse de celle rencontrée dans la classe précédente. Chez les expérimenté(e)s on trouve le critère "*Contrôle utilisateur*" et les critères suivants chez les inexpérimenté(e)s : "*Gr./Dist. par la localisation*", "*Concision*", "*Actions minimales*", "*Charge mentale*", "*Actions explicites*", "*Prise en compte de l'expérience de l'utilisateur*" et "*Correction des erreurs*".

Cette catégorie et la précédente regroupent au total quinze *critères élémentaires* sur 18. Chaque critère se trouve soit dans l'une, soit dans l'autre catégorie et sa position est commune ou non aux deux groupes. L'objectif étant de fournir des définitions de critères qui soient aussi précises que possible seuls les critères de ces deux catégories

Tableau 4

Classification des critères élémentaires selon que leur définition doit ou non subir des modifications

Sans modification	À modifier
Gr./Dist localisation	Prompting
Gr./Dist format	Clarté
Feed-back	Concision
Charge mentale	Actions minimales
Actions explicites	Contrôle utilisateur
Expérience utilisateur	Flexibilité
Protection contre erreurs	Homogénéité
Qualité messages	Signifiance des codes
Corrections des erreurs	Compatibilité

communs aux deux groupes ne feront l'objet d'aucune modification. Cette restriction exclue donc les critères "*Clarté*", "*Signifiance des codes*", "*Compatibilité*", "*Contrôle utilisateur*", "*Actions minimales*" et "*Concision*". Les critères qui peuvent être qualifiés de convenables du point de vue de leur définition et ce, pour les deux groupes, sont donc au nombre de neuf, soit la moitié de l'ensemble des *critères élémentaires*. Il s'agit donc des *critères élémentaires* suivants : "*Feedback*", "*Protection contre les erreurs*", "*Qualité des messages*" "*Gr./Dist. par le format*", "*Gr./Dist. par la localisation*", "*Charge mentale*", "*Actions explicites*", "*Prise en compte de l'expérience de l'utilisateur*" et "*Correction des erreurs*".

3.3.3. Critères élémentaires nécessitant une amélioration de leur définition

De l'ensemble des *critères élémentaires* neuf nécessitent des améliorations de leur définition, que ce soit pour les deux groupes ou un seul des deux. Il s'agit des critères "*Homogénéité*", "*Flexibilité*", "*Prompting*", "*Clarté*", "*Contrôle utilisateur*", "*Signifiance des codes*", "*Actions minimales*", "*Compatibilité*" et "*Concision*". Le tableau 4 résume ces données.

3.4. Confusions systématiques entre critères

Tous les critères nécessitant des améliorations sont confondus de façon systématique avec un ou plusieurs autres critères soit par l'un ou l'autre groupe ou soit par les deux. Toutefois, des confusions systématiques ont aussi été observées pour quatre des critères qualifiés de convenables. Ces quatre critères sont donc inclus dans la liste des critères dont les définitions devront être affinées.

Tableau 5

Confusions systématiques entre critères élémentaires chez les participant(e)s expérimenté(e)s et inexpérimenté(e)s (*)

Critères identifiés par les participant(e)s	Critères théoriques confondus	
	Expérimenté(e)s	**Inexpérimenté(e)s**
Prompting	Gr./Dist format (10^8) Feed-back immédiat (3^3) Concision (5^3) Protection c. erreurs (4^3) Compatibilité (5^5)	Gr./Dist format (4^3) Feed-back immédiat (6^5) Concision (3^3)
Gr./Dist. localisation		Signifiance codes (5^5)
Gr./Dist format		Prompting (3^3) Gr./Dist. localisation (5^4)
Feed-back immédiat		Contrôle utilisateur (3^3)
Clarté		Charge mentale (5^5)
Concision	Charge mentale (9^6)	Charge mentale (5^4)
Actions minimales	Concision (9^6) P. c. expérience (5^4)	Concision (8^6)
Actions explicites	Contrôle utilisateur (4^3)	
Contrôle utilisateur	Correction des erreurs (6^4)	Correction des erreurs (9^7) Actions minimales (5^4) Actions explicites (6^6)
Flexibilité	P. c. expérience (8^7) Contrôle utilisateur (4^3) Correction des erreurs (4^3)	P. c. expérience (11^8) Actions minimales (6^5)
Homogénéité	Gr./Dist. localisation (3^3)	Gr./Dist format (4^4)
Signif. codes		Concision (4^3)
Compatibilité	Concision (3^3) Flexibilité (4^4)	

(*) Les chiffres entre parenthèses indiquent les fréquences d'identification. L'exposant indique le nombre de participant(e)s qui y concoure. Les critères identifiés sont les réponses données par les participant(e)s au différents énoncés de problèmes. Les critères théoriques confondus sont les critères qui étaient associés à chacun des énoncés.

Le tableau 5 présente les critères qui seront améliorés de même que ceux avec lesquels ils sont confondus de façon systématique. Sur ce tableau on remarque par exemple que les expérimenté(e)s ont identifié un certain nombre de fois le *critère*

élémentaire "*Prompting*" alors que les énoncés se rapportaient théoriquement aux critères "*Gr./Dist. format*", "*Feed-back immédiat*", "*Concision*", "*Protection c. erreurs*" et "*Compatibilité*".

3.4.1. Confusions systématiques avec un seul autre critère élémentaire

Les deux groupes de participant(e)s se distinguent quant à leurs confusions systématiques uniques des *critères élémentaires*. Soit ces derniers font l'objet d'une confusion systématique unique pour un seul des deux groupes, soit ils font l'objet d'une confusion pour les deux groupes, auquel cas la confusion peut être ou non la même. Ainsi les critères "*Gr./Dist. par la localisation*", "*Feedback immédiat*", "*Clarté*", "*Actions minimales*" et "*Signifiance des codes*" identifiés par les participant(e)s ne sont confondus avec un seul autre que par les inexpérimenté(e)s. Chez les expérimenté(e)s seuls les critères "*Actions explicites*" et "*Contrôle utilisateur*" font l'objet d'une confusion systématique unique. Le critère "*Homogénéité*" entraîne des confusions uniques différentes chez les deux groupes. Quant à la "*Concision*", elle entraîne une confusion systématique identique chez les deux groupes.

Ces résultats indiquent que la difficulté que posent les définitions de ces *critères élémentaires* ne semblent pas pouvoir se réduire à leur trop grande généralité. Il s'agit davantage d'une insuffisance de leur pouvoir discriminatif d'une part et d'un manque de saillance d'autre part, ces dernières ne produisant pas les mêmes performances et les mêmes erreurs chez les deux groupes. Ce manque de saillance laisse donc place à d'autres significations pouvant différer selon l'expérience. Les choses semblent différentes pour les critères entraînant des confusions systématiques avec plusieurs autres critères.

3.4.2. Confusions systématiques avec plusieurs autres critères élémentaires

Six *critères élémentaires* entraînent des confusions systématiques avec plusieurs autres critères. Il s'agit : du "*Prompting*", du "*Gr./Dist. par le format*", des "*Actions minimales*", du "*Contrôle utilisateur*", de la "*Flexibilité*" et de la "*Compatibilité*". Deux de ces critères présentent ce type de confusions pour les deux groupes ("*Prompting*" et "*Flexibilité*"), deux pour les inexpérimenté(e)s ("*Gr./Dist. par le format*", "*Contrôle utilisateur*") et deux pour les expérimenté(e)s ("*Actions minimales*" et "*Compatibilité*").

Ces critères, contrairement à ceux qui produisent des confusions uniques, semblent être définis de manière trop générale, ce qui a pour effet d'entraîner des confusions systématiques avec plusieurs autres *critères élémentaires*.

Tableau 6

Fréquences des identifications incorrectes pour chacun des énoncés dans le cas des confusions systématiques et probabilités associées aux tests binômiaux.

Critères identifiés	Numéro de l'énoncé	Critères théoriques	Participant(e)s		Total	Prob. associées au test binômial
			Expéri-menté(e)s	Inexpéri-menté(e)s		
Prompting	29	Gr./Dist. format	7	2	9	0,212
	54		3	2	5	
	23	Feed-back immédiat	2	5	7	0,090
	46		1	1	2	
	12	Concision	2	0	2	0,145
	32		3	3	6	
Concision	1	Charge mentale	4	3	7	0,605
	8		5	2	7	
Actions minimales	12	Concision	4	3	7	0,315
	32		5	5	10	
Contrôle utilisateur	45	Corr. des erreurs	4	6	10	0,151
	58		2	3	5	
Flexibilité	17	Expér. utilisateur	5	6	11	0,324
	57		3	5	8	

Pour pouvoir utiliser les confusions systématiques, qu'elles soient uniques ou multiples, lors des modifications de définitions, il faut d'abord s'assurer que ces confusions résultent bien des définitions et non pas des énoncés de problèmes. Étant donné qu'il y a deux énoncés de problème par *critère élémentaire*, on est en droit de s'attendre à ce que les fréquences des erreurs ne diffèrent pas de façon significative d'un énoncé à l'autre dans le cas où les définitions seraient en cause. Si les énoncés sont responsables des erreurs d'identifications, alors des fréquences différentes pour l'un et l'autre des énoncés devraient s'observer.

Le tableau 6 présente les fréquences d'identifications incorrectes pour chacun des énoncés se rapportant aux critères ayant fait l'objet de confusions systématiques identiques chez les deux groupes. Les tests binômiaux qui ont été effectués pour déterminer si le total des fréquences diffère d'un énoncé à l'autre indiquent qu'aucune des différences n'atteint le seuil de signification de 0,05. On peut donc conclure que les confusions systématiques résultent plutôt des définitions de critères que des énoncés de

problèmes. Néanmoins, ces derniers pourront être pris en compte lors des modifications de définitions.

3.5. Exemples d'analyse des confusions systématiques pour modifications ultérieures des définitions

Les modifications des définitions s'appuient sur les confusions systématiques évoquées dans les paragraphes précédents. Deux exemples seront présentés.

Le premier exemple concerne le "*Prompting*". Ce dernier est le *critère élémentaire* qui pose le plus de problèmes. Il a été confondu systématiquement avec plusieurs autres critères. Parmi ceux-ci, trois ont été confondus par au moins cinq participant(e)s. Chez les expérimenté(e)s huit l'ont confondu avec le "*Gr./Dist. par le format*" et cinq avec la "*Compatibilité*". Chez les inexpérimenté(e)s, cinq l'ont confondu avec le "*Feedback immédiat*". De plus, les critères "*Gr./Dist. par le format*" et "*Feedback immédiat*" ont été confondus systématiquement par les deux groupes.

Ces deux dernières confusions concernent des *critères élémentaires* voisins, i.e. appartenant au même *critère principal*. On remarque par ailleurs que chez les inexpérimenté(e)s le critère "*Gr./Dist. par le format*" est identifié à la place des critères "*Prompting*" et "*Gr./Dist. par la localisation*". Réciproquement, les énoncés relatifs au "*Prompting*" se voient associer le critère "*Gr./Dist. par le format*". Ces *critères élémentaires* semblent donc difficiles à distinguer les uns des autres.

La définition présentée aux participant(e)s était la suivante :

> "*Le prompting correspond aux informations fournies à l'utilisateur, relatives : à l'état dans lequel celui-ci se trouve ; aux actions possibles ou attendues et aux moyens de les mettre en œuvre ; aux aides disponibles ; et aux formats d'entrée de données*".

Le "*Prompting*" a été identifié, de façon systématique, pour les énoncés se rapportant au "*Gr./Dist. par le format*". Le premier énoncé est le suivant (voir l'illustration à la figure 2) :

> "*Lorsque plusieurs options apparaissent sur un écran (e.g.,* **"Dict."**, **"Annuler"**, **"OK"** *écran 2), rien n'indique l'option par défaut. Sur l'écran 2 l'option* **"OK"** *est l'option par défaut et elle peut être sélectionnée par un simple retour de chariot*".

Théoriquement, il s'agit ici de l'absence de distinction, par le format, entre items. Trois options disponibles, i.e. pouvant être sélectionnées, sont présentées à l'utilisateur. Une d'entre elles est l'option par défaut mais rien ne permet de la distinguer des deux autres. Or, généralement ce type d'options est distingué par un contour gras (Il s'agit

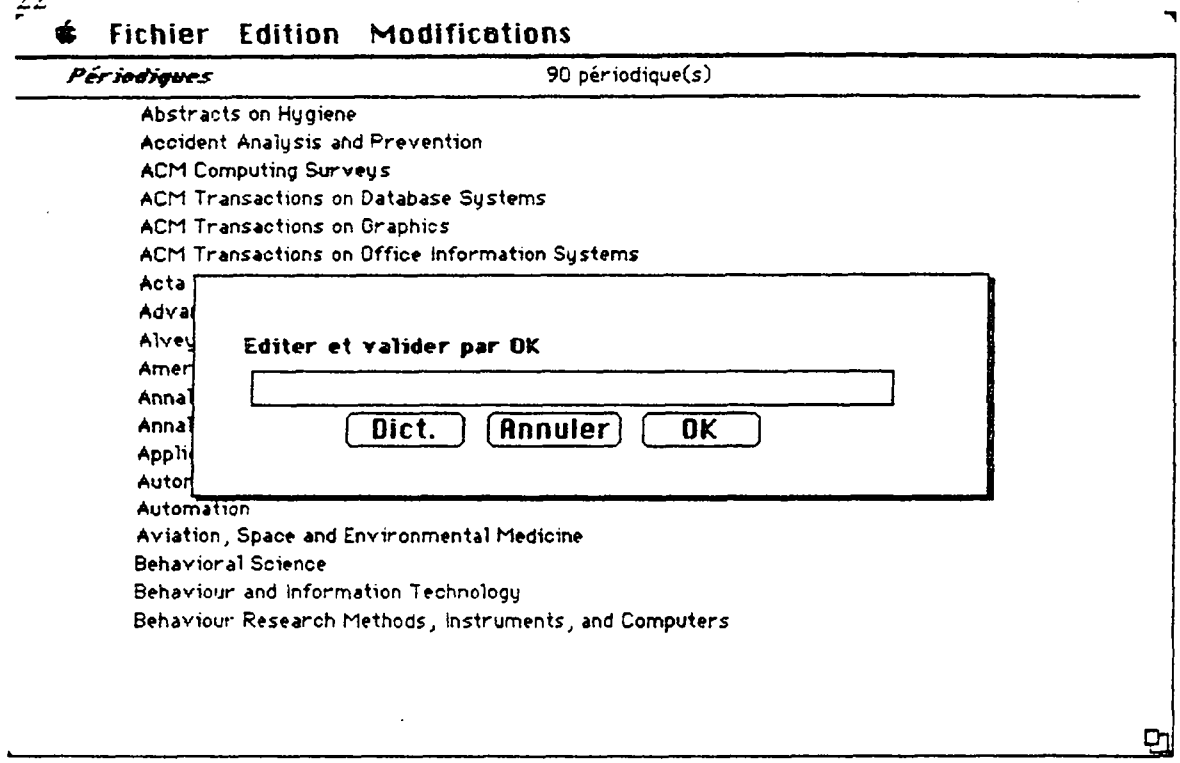


Figure 2. Copie d'écran illustrant l'énoncé du problème numéro 54 se rapportant au critère élémentaire "Groupement / Distinction par le format".

d'ailleurs d'un standard Microsoft, 1987). Par ailleurs, on indique dans l'énoncé du problème qu'un équivalent clavier (touche retour de chariot) est associé à cette option. L'absence de distinction et l'existence d'un équivalent clavier associé à une option sans qu'il soit possible de savoir à laquelle, peut être une source d'erreur et à plus forte raison si d'autres associations entre équivalents claviers existent. Les associations ne sont pas toujours les mêmes. Ainsi, si dans certains cas l'option par défaut est l'option "**OK**", dans d'autres il peut s'agir de l'option "**Annuler**". Distinguer les options par défaut des autres permet à l'utilisateur de sélectionner l'option choisie et d'effectuer les actions appropriées plus rapidement. Dans cet exemple l'utilisateur se voit présenter les options correspondant aux actions possibles. Il sait par conséquent ce qu'il peut faire dans ce contexte. Il n'y a donc pas de problème de "*Prompting*" mais bien un problème de distinction entre items. L'erreur d'identification semble résulter d'une confusion entre présentation et distinction des options.

Le deuxième énoncé était le suivant :

"*Dans la fenêtre de l'écran 4 apparaissent les labels* "**Titre**" *et* "**REF**". *La case située sous le label* "**Titre**" *laisse supposer qu'il s'agit d'une aire de saisie tout comme l'aire située sous* "**REF**". *Il s'agit en fait d'une aire de message dans laquelle s'affiche le titre associé à la* "**REF**" (voir figure 3).

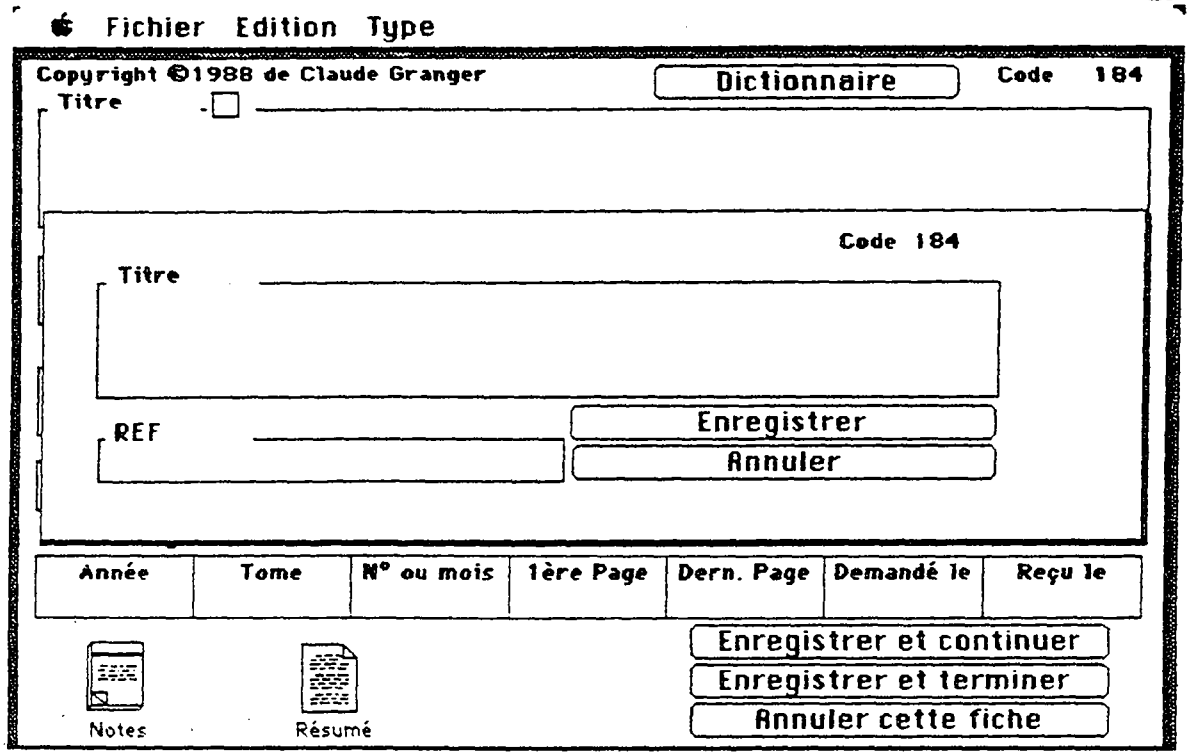


Figure 3. Copie d'écran illustrant l'énoncé du problème numéro 29 se rapportant au critère élémentaire "Groupement / Distinction par le format".

Dans cet exemple deux aires identifiées par des labels différents sont délimitées par des contours identiques bien qu'il s'agisse dans un cas d'une aire de message et dans l'autre d'une aire de saisie. Bien que l'énoncé du problème distingue l'aire de saisie de l'aire de message on ne sait pas si l'utilisateur sait qu'il s'agit d'aires de nature différente. Si le/la participant(e) fait l'hypothèse que l'utilisateur sait qu'il s'agit de deux aires différentes alors il répondra en identifiant le critère "*Gr./Dist. par le format*". Le/la participant(e) peut faire l'hypothèse dans ce cas que l'utilisateur sait ce qu'il doit faire bien que les deux aires ne soientt pas distinctes. Toutefois si le/la participant(e) fait l'hypothèse que l'utilisateur ne connaît pas cette distinction et qu'il ne sait pas qu'il doit saisir l'information sous "**REF**" alors il répondra "*Prompting*". Ces deux points de vue sur ce problème expliquent probablement les réponses données. Dans ce cas-ci, les erreurs d'identification semblent pouvoir s'expliquer en partie par l'énoncé du problème. Une solution à apporter à cet énoncé afin qu'il ne se rapporte qu'à un seul critère serait de placer un curseur dans l'aire de saisie située sous "**REF**".

A partir de ces deux exemples nous savons que la définition du "*Prompting*" devra préciser l'idée que le guidage se fait au niveau des informations sur les actions attendues et permises de l'utilisateur plutôt que sur l'organisation de ces informations. Ainsi par exemple, dans le premier cas, on ne présente à l'utilisateur que les options

disponibles de façon à ce qu'il ne puisse pas en sélectionner qui soient sans effet. Dans le deuxième cas ces options disponibles sont organisées de façon à ce que leur recherche et leur sélection soient plus faciles. La modification de la définition de ce critère devra aussi tenir compte des autres confusions systématiques.

Le deuxième exemple concerne le critère "*Concision*". Ce *critère élémentaire* a été confondu avec un seul autre critère, la "*Charge mentale*" et ce par les deux groupes de participant(e)s. Ici aussi les confusions apparaissent au sein d'un même *critère principal*, à savoir la "*Charge de travail*".

La définition de la "*Concision*" donnée aux participant(e)s était la suivante :

"*La concision concerne la charge de travail au niveau perceptif et mnésique en rapport à des éléments individuels d'entrée ou de sortie*".

La "*Charge mentale*" pour sa part était définie comme suit :

"*La charge mentale concerne la charge de travail du point de vue perceptif et mnésique, pour des ensembles d'éléments*".

Ce qui semble avoir posé problème pour ces énoncés, chez quelques participant(e)s, est la distinction entre éléments individuels et ensemble d'éléments. Par ailleurs la "*Concision*" concerne les entrées de données alors que la "*Charge mentale*" concerne davantage les affichages. Ces points devront apparaître plus explicitement dans les nouvelles définitions.

On notera au passage que toutes les améliorations pouvant être apportées aux définitions des *critères élémentaires* ne représentent pas la même importance. En effet, les définitions des critères faisant l'objet de *confusions systématiques* mutiples nécessitent probablement plus de travail que les définitions des critères ne faisant l'objet que de confusions uniques. On doit donc se garder d'une part de juger de la même manière tous les critères présentant des *confusions systématiques* et à plus forte raison si ces confusions ne sont propres qu'à un seul des deux groupes.

4. DISCUSSION

L'objectif principal de cette recherche était de valider des critères ergonomiques pour l'évaluation des interfaces utilisateurs. Plus précisément, il s'agissait de savoir si ces critères, en fonction de leurs définitions, justifications et exemples, pouvaient permettre à des participant(e)s ayant une formation et une expérience différentes, d'étiqueter de façon univoque des erreurs de conception. La tâche choisie à cette fin consistait à fournir à des participant(e)s expérimenté(e)s et inexpérimenté(e)s des

"erreurs", d'un point de vue ergonomique, de conception d'une interface utilisateur et à leur demander d'identifier le *critère élémentaire* qui leur semblait mis en cause.

Les résultats obtenus sont intéressants à plusieurs égards. Ils permettent de tirer des conclusions importantes quant à la lisibilité des définitions, justifications et exemples fournis, quant aux performances globales à la tâche d'identification des critères et finalement sur la qualité globale des définitions. Toutes ces données permettent de juger la validité des critères, définitions et exemples proposés.

Les résultats de cette recherche indiquent premièrement que l'expérience des participant(e)s n'influence pas le temps nécessaire à la lecture des définitions, des exemples et des justifications qui accompagnent chaque critère. Les termes choisis pour les définitions, les justifications et les exemples ne semblent poser de problèmes ni pour l'un ni pour l'autre groupe. On peut donc en déduire que le degré de lisibilité, dans la mesure où il peut être évalué à partir de la rapidité de lecture, semble approprié aux participant(e)s (Hartley, 1990). De plus, l'expérience n'a d'influence ni sur le temps consacré à la tâche d'identification des *critères élémentaires*, ni sur le nombre d'identifications correctes et ce, que ce nombre soit calculé à partir des *critères élémentaires* ou des *critères principaux*. Ces deux modes de calcul montrent par ailleurs que si certaines erreurs sont dues à des problèmes de discrimination entre *critères élémentaires*, d'autres résultent de confusions entre *critères principaux*.

Bien que les résultats précédents ne diffèrent pas d'un groupe à l'autre, les fréquences moyennes d'identification et les proportions d'identifications correctes des *critères élémentaires* quant à elles diffèrent. Les trois classes de critères (critères bien définis, définis de façon satisfaisante et nécessitant des améliorations) établies sur la base de ces indices contiennent des *critères élémentaires* différents selon le groupe. Les expérimenté(e)s et les inexpérimenté(e)s n'utilisent pas aussi souvent et avec la même exactitude tous les *critères élémentaires*. L'expérience et/ou la formation des participant(e)s semblent donc interférer avec certaines des définitions de critères données par les expérimentateurs. En d'autres termes, le contrôle qu'exercent certaines définitions actuelles sur les identifications des participant(e)s n'arrive pas à se substituer à celui des définitions qu'évoquent ces critères. Si les définitions permettaient un bon apprentissage discriminatif les participant(e)s répondraient de la même façon, i.e. identifieraient les mêmes *critères élémentaires* en présence d'exemples d'erreurs de conception. Certains critères actuels ne semblent pas avoir, a priori, forcément la même signification pour tous les participant(e)s. Les significations accordées à ces termes ou leur interprétation reflètent probablement la pratique ou l'expérience de ces derniers. Ces significations peuvent donc ne pas correspondre à celles qu'on voudraient bien leur

donner ici. Les définitions actuelles des critères disponibles dans la littérature et plus particulièrement à leur manque d'homogénéité contribue sans doute à cette situation.

Les travaux de Gillan et Breedin (1990) corroborent en partie l'idée suivant laquelle l'expérience modifie la signification des termes d'un domaine. Ces auteurs remarquent que des spécialistes du facteur humain, ayant une expérience en conception d'interface, des développeurs et des utilisateurs de logiciels ont des connaissances déclaratives différentes du domaine de l'IPO. Cela se reflètent par une organisation, une classification et des liens entre des termes se rapportant aux interfaces utilisateurs qui varient selon les groupes. On peut donc supposer qu'il en est de même pour les critères.

Les classes de *critères élémentaires* ont permis de déterminer ceux qui bénéficieraient d'une amélioration de leur définition. Ces *critères élémentaires* sont au nombre de 13 sur 18. Ce nombre doit cependant être commenté. La sélection des critères devant subir des améliorations s'est effectuée d'une façon stricte : seuls les critères bien définis et définis de façon satisfaisante pour les deux groupes n'entraîneront pas d'améliorations. De plus, toutes les définitions ne nécessitent pas le même travail. Certaines exigent des modifications très légères et très précises : on trouve dans cette catégorie les *critères élémentaires* faisant l'objet d'une *confusion systématique* avec un seul autre critère ; et d'autres exigeant des modifications plus importantes, comme c'est le cas pour les *critères élémentaires* à *confusions systématiques* multiples. Pour guider ces améliorations les matrices de confusions fournissent des indications fort utiles. Elles permettent de déceler les confusions systématiques entre *critères élémentaires* (Bakeman et Gottman, 1986). Sachant quels sont les critères confondus les uns avec les autres, il n'y a par la suite qu'à se pencher davantage sur les termes utilisés dans les définitions. Il est possible que ces confusions résultent de termes voisins.

Le fait d'avoir à modifier un certain nombre de définitions ne doit pas nous faire perdre de vue l'idée que les résultats obtenus dans cette recherche sont encourageants. Rappelons que les conditions dans lesquelles se déroule la tâche sont assez difficiles. Dans la première phase de l'expérience, les participant(e)s avaient accès au document présentant les définitions, les justifications et les exemples. Dans la deuxième phase, ils/elles n'avaient accès qu'aux seules définitions. Les participant(e)s disposaient donc de très peu d'informations pour identifier les *critères élémentaires*. Dans l'état actuel des définitions les coefficients *Kappa* pour les *critères élémentaires* sont quand même supérieurs à 0.50 et les scores globaux varient de 56 à 77,6% selon qu'il s'agit des *critères élémentaires* ou des *critères principaux*.

A ces conditions de passation s'ajoute la brièveté de la phase d'apprentissage, qu'on devrait peut être plus justement qualifier de phase de lecture. L'acquisition de concepts nouveaux n'est pas une tâche simple et à plus forte raison si ces concepts sont proches d'autres concepts déjà connus, pas plus d'ailleurs que ne l'est celle qui consiste à identifier un concept à partir d'un exemple s'y rapportant. Des auteurs tels que Gropper (1983) et Johnson et Chase (1981) par exemple, ont identifié les exigences auxquelles on doit satisfaire pour assurer la compréhension d'un concept. Tout d'abord, une définition du concept doit être apprise. Puis, pour permettre une bonne *discrimination* entre concepts, les exemples propres à chacun doivent être le plus distinct possible des exemples d'un autre concept. Par ailleurs, si l'on veut favoriser la *généralisation* du concept à des exemples non enseignés chaque concept doit être accompagné d'un certain nombre d'exemples représentatifs de la classe mais aussi différents que possible les uns des autres (voir aussi Richard, 1990). Pour faciliter ces deux processus on doit donc fournir, pour chaque concept, des exemples et des contre-exemples.

Lorsqu'un concept est bien appris les faits s'y rapportant peuvent être énoncés et les liens rattachant les différents concepts évoqués. Par ailleurs leur contenu peut être présenté sous des formes différentes, des exemples réels ou concrets peuvent être identifiés et de nouveaux fournis. L'extension générique d'un concept permet à un individu d'évoquer un concept qu'évoqueraient des experts face à une situation, ou à un objet nouveau (Johnson et Chase, 1981).

Or l'acquisition des concepts nécessite un feedback constant lors de l'apprentissage et ce d'autant plus que les concepts sont difficiles à expliciter. Il faut, lors de cet apprentissage, identifier et corriger les généralisations erronées que peuvent faire les individus. Dans la phase d'apprentissage de cette recherche, aucune interaction ne survient entre l'expérimentateur et le(s) participant(e)s. Lorsque ces dernier(e)s indiquent à l'expérimentateur qu'ils ont terminé la lecture du premier document celui-ci fait l'hypothèse que les critères, leur définition, leurs justifications et exemples ont été compris. Cette méthodologie suppose plus ou moins que l'apprentissage ne se poursuit pas au-delà de cette phase. Or il est probable que les participant(e)s modifient l'idée qu'ils/elles se font des critères sur la base des réponses qu'ils/elles fournissent à chacun des problèmes. L'absence d'information rétroactive sur l'exactitude d'une réponse, ou si l'on veut le silence des expérimentateurs, est souvent interprétée comme s'il s'agissait d'un acquiescement (voir George, 1983). Il est donc possible que les participant(e)s modifient ou interprètent différemment les définitions des critères à partir de leurs premières réponses. D'un point de vue formation il serait peut-être

avantageux de fournir aux participant(e)s une connaissance de leurs résultats lors de la tâche d'identification tout comme il serait pédagogique de leur fournir une phase d'apprentissage plus interactive de façon à pouvoir évaluer leur compréhension des concepts tout au long de l'acquisition. Cette dernière peut modifier de façon durable les connaissances que possède un individu (George, 1983) tout en lui permettant d'affiner la compréhension qu'il a des définitions de critères et des exemples concrets auxquels ils se rapportent.

Ces divers points font apparaître l'écart pouvant exister entre ce qui pourrait être un enseignement optimal des concepts que représentent les critères ergonomiques pour l'évaluation des interfaces utilisateurs et la tâche dans laquelle ont été placé(e)s les participant(e)s. Cet écart nous fait penser que les résultats obtenus sont fort encourageants. Des améliorations de certaines définitions devraient permettre une meilleure identification des critères à partir d'exemples.

Les résultats de cette recherche montrent que la définition univoque de critères n'est pas une entreprise facile. Contrairement à ce que l'on pourrait croire, les critères ergonomiques qu'on trouve ici et là dans la littérature peuvent recouvrir des significations fort différentes selon l'expérience de ceux qui les utilisent. Une mise en garde contre l'utilisation de critères plus ou moins définis s'impose donc et à plus forte raison lorsque ces derniers sont utilisés comme outil de recherche. Il n'est pas rare de voir des critères ergonomiques utilisés pour la classification des problèmes détectés dans les études sur les évaluations d'interfaces utilisateurs par des ergonomes. Or la plupart du temps ces critères, empruntés à diverses sources ou reflétant l'expérience des expérimentateurs, ne sont pas explicitement définis ou alors ils le sont très succinctement. Par ailleurs et, ce qui est peut être plus grave encore, les résultats de ces classifications ne font l'objet d'aucun test explicite d'accord inter-juge (ex.: Molich et Nielsen, 1990). Ce dernier point est extrêmement important si l'on se rappelle que les douze participant(e)s expérimenté(e)s de la présente étude ont obtenu un coefficient *Kappa* de 0,61. Il serait difficile de publier une recherche dont les conclusions se basent sur des données issues d'une classification à propos de laquelle l'accord inter-juge n'est pas plus élevé. Ce problème de la reproductibilité des données se pose dès qu'il s'agit d'utiliser des grilles d'analyse, d'observation ou de classification. Lorsque les accords inter-juges ne sont pas suffisamment élevés, diverses solutions peuvent être envisagées (voir par exemple l'étude de Baril-Gingras et Lortie, 1989). Une solution peut consister à fournir aux participant(e)s une formation plus poussée et une autre à modifier les définitions des catégories. Compte tenu des objectifs poursuivis, la modification des définitions et éventuellement des justifications et des exemples s'avère être la stratégie la plus adaptée. Après modification, les définitions devront subir une autre validation.

Lorsque les définitions de critères auront été affinées et re-validées, la tâche subséquente consistera à évaluer l'aide qu'un tel jeu de critères peut apporter à des spécialistes et non-spécialistes de l'ergonomie des interfaces dans leur tâche d'évaluation. Déjà certaines difficultés auxquelles il faudra faire face peuvent être envisagées. Notons entre autres le problème de l'application des critères aux différents objets de l'interface, l'importance relative des critères selon la tâche ou encore les liens entre ces critères et les différentes mesures de performance des utilisateurs. Ces problèmes sont de taille car il existe à l'heure actuelle très peu de données sur ces questions. Si, par exemple, certaines recherches ont pu mettre en évidence les effets de l'"*Homogénéité*" sur divers indices de performance (voir à ce sujet Polson, 1988), les données sont beaucoup plus rares quant aux autres critères. Ces questions sont cependant fondamentales pour la mise au point d'une méthodologie d'évaluation basée sur des critères ergonomiques.

RÉFÉRENCES

Bakeman, R. et Gottman, J. M. (1986). *Observing interaction : an introduction to sequential analysis*. Cambridge : Cambridge University Press.

Baril-Gingras, G. et Lortie, M. (1990). *L'observation en ergonomie : comment s'assurer de la reproductibilité des données ?* Comptes rendus du XXVIème congrès de la SELF (pp. 23-26). Montréal, 3-5 octobre.

Bellotti, V. (1988). Implications of current design practice for the use of HCI techniques. In D. M. Jones et R. Winder (Eds.), *People and computers IV* (pp. 13-34). Cambridge : Cambridge University Press.

Brown, C. M. (1988). *Human-computer interface design guidelines*. Norwood, NJ : Ablex.

Card, S. K., Moran, T. P. et Newell, A. (1983). *The psychology of human-computer interaction*. London : Lawrence Erlbaum.

Caverni, J.-P. (1988). La verbalisation comme source d'observables pour l'étude du fonctionnement cognitif. In J.-P. Caverni, C. Bastien, P. Mendelsohn et G. Tiberghien (Eds.), *Psychologie cognitive. Modèles et méthodes* (pp. 253-273). Grenoble : Presses Universitaires de Grenoble.

Christie, B. et Gardiner, M. M. (1990). Evaluation of the human-computer interface. In J. R. Wilson et E. N. Corlett (Eds.), *Evaluation of human work : a practical ergonomics methodology* (pp. 271-320). London : Taylor et Francis.

Clegg, C. W., Warr, P. B., Green, T. R. G., Monk, A., Kemp, N., Allison, G. et Lansdale, M. (1988). *People and computers - how to evaluate your company's new technology*. Chichester : Ellis Horwood.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Gardiner, M. M. et Christie, B. (Eds.) (1987). *Applying cognitive psychology to user-interface design*. Chichester : Wiley et Sons.

Gardner, A., Mayfield, T. F. et Maguire, M. C. (1985). Human factors guidelines for the design of computer-based systems. In B. Shackel (Ed.), *Human-computer interaction - interact '84* (pp. 649-654). Amsterdam : Elsevier.

George, C. (1983). *Apprendre par l'action*. Paris : Presses Universitaires de France.

Gillan, D. J. et Breedin, S. D. (1990). Designers' models of the human-computer interface. In J. Carrasco et J. Whiteside (Eds.), *Empowering people : CHI '90 conference proceedings* (pp. 391-398). Reading, MA : Addison-Wesley.

Green, T. R. G., Schiele, F. et Payne, S. J. (1988). Formalisable models of user knowledge in human-computer interaction. In G. C. van der Veer, T. R. S. Green, J.-M. Hoc et D. M. Murray (Eds.), *Working with computers : theory versus outcome* (pp. 3-46). London : Academic Press.

Gropper, G. L. (1983). A behavioral approach to instructional prescription. In C. M. Reigeluth (Ed.), *Instructional-design theories and models : an overview of their current status* (pp. 101-161). Hillsdale, NJ : Lawrence Erlbaum.

Hammond, N. V., Hinton, G., Barnard, P. J., MacLean, A., Long, J. B. et Whitefield, A. (1985). Evaluating the interface of a document processor : a comparison of expert judgement and user observation. In B. Shackel (Ed.), *Human-Computer Interaction - Interact '84* (pp.725-730). Amsterdam : North-Holland.

Hartley, J. (1990). Is this chapter any use ? Methods for evaluating text. In J. R. Wilson et E. N. Corlett (Eds.), *Evaluation of human work* (pp. 248-270). London : Taylor et Francis.

Hayes, S. C. (1986). The case of the silent dog - verbal reports and the analysis of rules : a review of Ericsson and Simon's protocol analysis : verbal report as data. *Journal of Experimental Analysis of Behavior, 45*, 351-363.

Heckel, P. (1984). *The elements of friendly software design.* New York : Warner books.

Hoc, J.-M. et Leplat, J. (1983). Evaluation of different modalities of verbalizations in a sorting task. *International Journal of Man-Machine Studies, 18*, 283-306.

Johnson, K. R. et Chase, P. N. (1981). Behavior analysis in instructional design : a functional typology of verbal tasks. *The Behavior Analyst, 4(2)*, 103-121.

Johnson, G. I., Clegg, C. W. et Ravden, S. J. (1989). Towards a practical method of user interface evaluation. *Applied Ergonomics, 20(4)*, 255-260.

Karat, J. (1988). Software evaluation methodologies. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 891-903). Amsterdam : Elsevier.

Kieras, D. E. et Polson, P. G. (1985). An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies, 22*, 365-394.

Knowles, C. (1988). Can cognitive complexity theory (CCT) produce an adequate measure of system usability ? In D. M. Jones et R. Winder (Eds.), *People and computer IV* (pp. 291-307). Cambridge : Cambridge University Press.

Maguire, M. et Sweeney, M. (1989). System monitoring : garbage generator or basis for comprehensive evaluation system ? In A. Sutcliffe et L. Macaulay (Eds.), *People and computer V* (pp.375-394). Cambridge : Cambridge University Press.

Marshall, C., Nelson, C. et Gardiner, M. M. (1987). Design guidelines. In M. M. Gardiner et B. Christie (Eds.), *Applying Cognitive Psychology to User-Interface Design* (pp. 221-278). New York : John Wiley et Sons.

Microsoft (1987). *Microsoft Windows. Version 2.0.* Microsoft Corporation.

Molich, R. et Nielsen, J. (1990). Improving a human computer dialogue. *Communications of the ACM, 33/3*, 338-348.

Moran, T. P. (1981). The command language grammar : a representation of the user interface of interactive computer systems. *International Journal of Man-Machine Studies, 15*, 3-50.

Mosier, J. N. et Smith, S. L. (1985). Applications of guidelines for designing user interface software. *Proceedings of the Human Factors Society 29th Annual Meeting, 29*, 946-949.

Mosier, J. N. et Smith, S. L. (1986). Applications of guidelines for designing user interface software. *Behaviour and Information Technology, 5(1)*, 39-46.

Nielsen, J. et Molich, R. (1990). Heuristic evaluation of user interfaces. In J. Carrasco et J. Whiteside (Eds), *Empowering People. CHI'90 Conference Proceedings* (pp. 249-256). Reading, MA : Addison-Wesley.

Oppermann, R., Murchner, B., Paetau, M., Simm, H. et Stellmacher, I. (1989). *Evaluation of dialog systems.* (GMD-Studien Nr. 169). Sankt Augustin, Allemagne : Gesellschaft Für Mathematik Und Datenverarbeitung MBH.

Pollier, A. (1991). *Evaluation d'une interface par des ergonomes : diagnostics et stratégies* (Rapport de recherche n° 1391). Rocquencourt, France : Institut National de Recherche en Informatique et en Automatique.

Polson, P. G. (1988).The consequences of consistent and inconsistent user interfaces. In R. Guindon (Ed.), *Cognitive science and its applications for human-computer interaction* (pp.59-108). Hillsdale, New Jersey : Erlbaum.

Ravden, S. J. (1988). Ergonomic criteria for design of the software interface between human and computer in CIM. *International Journal of Computer Applications in Technology, 1*, 35-42.

Ravden, S. J., et Johnson, G. I. (1989). *Evaluating usability of human-computer interfaces : a practical method. Chichester* : John Wiley et Sons.

Reisner, P. (1983). Formal grammars as a tool for analysing ease of use : some fundamental concepts. In J. C. Thomas et M. Schneider (Eds.), *Human factors in computing systems.* Norwood, NJ : Ablex.

Richard, J. F. (1990). *Les activités mentales : comprendre, raisonner, trouver des solutions.* Paris : Armand Colin.

Rivlin, C., Lewis, R. et Cooper, R. D. (1990). *Guidelines for screen design.* Oxford : Blackwell Scientific Publications.

Rubinstein, R. et Hersh, H. (1984). *The human factor : designing computer system for people.* Burlington, MA : Digital press.

Scapin, D. L. (1986). *Guide ergonomique de conception des interfaces homme-machine* (Rapport technique n° 77). Rocquencourt, France : Institut National de Recherche en Informatique et en Automatique.

Scapin, D. L. (1990a). Des critères ergonomiques pour l'évaluation et la conception d'interfaces utilisateurs. *Actes du XXVI Congrès de la SELF*, 3-5 Octobre, Montréal, Canada.

Scapin, D. L. (1990b). Organizing human factors knowledge for the evaluation and design of interfaces. *International Journal of Human-Computer Interaction, 2(3)*, 203-229.

Senach, B. (1990). *Evaluation ergonomique des interfaces homme-machine : une revue de la littérature.* (Rapport de recherche n° 1180). Sophia Antipolis, France : Institut National de Recherche en Informatique et en Automatique.

Schneiderman, B. (1987). *Designing the user interface : strategies for effective human-computer interaction.* Reading, MA : Addison-Wesley.

Smith, S. L. (1988). Standards versus guidelines for designing user interface software. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 877-889). Amsterdam : Elsevier.

Smith, S. L. et Mosier, J. N. (1986). *Guidelines for designing user interface software* (Rapport ESD-TR-86-278). Bedford, Massachusetts : Mitre.

ANNEXES

Annexe 1

Définitions, justification et exemples associés aux critères

Critères

1. Guidage

Définition :

Le guidage est l'ensemble des moyens mis en œuvre pour conseiller, orienter, informer, et conduire l'utilisateur lors de ses interactions avec l'ordinateur (messages, alarmes, labels, etc.). Quatre sous-critères participent au guidage : le "**Prompting**", les "**Groupements et distinctions entre items**", le "**Feed-back immédiat**", et la "**Clarté**".

Justification(s) :

Un bon guidage facilite l'apprentissage et l'utilisation du système en permettant à l'utilisateur : de savoir, à tout moment, où il se trouve dans une séquence d'interactions, ou dans l'accomplissement d'une tâche ; de connaître les actions permises ainsi que leurs conséquences ; et d'obtenir de l'information supplémentaire sur demande. La facilité d'apprentissage et d'utilisation qui s'ensuive permet de meilleures performances et occasionne moins d'erreurs.

1.1. Prompting

Définition :

Le prompting correspond aux informations fournies à l'utilisateur, relatives : à l'état dans lequel celui-ci se trouve ; aux actions possibles ou attendues et aux moyens de les mettre en œuvre ; aux aides disponibles ; et aux formats d'entrée de données.

Justification(s) :

Un bon prompting guide l'utilisateur et lui évite par exemple d'avoir à apprendre une série de commandes. Il permet aussi à l'utilisateur de savoir dans quel mode il est et où il se trouve dans le dialogue, et ce qu'il a fait pour s'y trouver. Un bon prompting aide donc l'utilisateur à se déplacer dans un système et à savoir les opérations qu'il peut effectuer.

Exemples de recommandations :

- Guider les entrées de données par un format adéquat.
- Indiquer des unités de mesure.
- Indiquer toutes les informations d'état.

1.2. Groupement/Distinction entre items

1.2.1. Groupement/Distinction par la localisation

Définition :

Le critère Groupement/Distinction par la localisation se réfère plus particulièrement au positionnement des items les uns par rapport aux autres dans le but d'indiquer leur appartenance ou non-appartenance à une même classe, ou encore dans le but de montrer la distinction entre différentes classes.

Justification(s) :

La compréhension d'un écran dépend, entre autres choses, de l'arrangement des objets, images, textes, et commandes présentés sur un écran. L'utilisateur aura plus de facilité à repérer les différents items s'ils sont organisés logiquement. De même, il pourra mieux les apprendre et s'en rappeler.

Exemples de recommandations :

- Organisation des items en listes hiérarchiques.
- Grouper les options de menus en fonction des objets sur lesquels elles s'appliquent.

1.2.2. Groupement/Distinction par le format

Définition :

Le critère Groupement/Distinction par le format se réfère plus particulièrement aux indices graphiques (format, couleur, etc.) permettant de faire apparaître l'appartenance ou la non-appartenance d'items à une même classe, ou encore permettant d'indiquer la distinction entre différentes classes.

Justification(s) :

L'utilisateur aura plus de facilité à connaître les liens entre items ou classes d'items si des formats, ou codages permettent d'illustrer leurs relations de proximité et/ou de différence. De même, il pourra mieux les apprendre et s'en rappeler.

Exemples de recommandations :

- Etablir une distinction entre des aires ayant des fonctions différentes (commande, message, etc.).
- Etablir une distinction entre les labels et les champs d'entrée.

1.3. Feed-back immédiat

Définition :

Le Feed-back immédiat se réfère aux réponses de l'ordinateur consécutives aux actions de l'utilisateur. L'ordinateur doit répondre à toutes les actions de l'utilisateur, le plus rapidement possible.

Justification(s) :

La qualité et la rapidité du feed-back sont deux facteurs importants : pour l'établissement de la confiance et de la satisfaction de l'utilisateur ; pour la compréhension du dialogue. Ces facteurs permettent à l'utilisateur de se faire une bonne représentation du système. Des réponses lentes de l'ordinateur entraînent souvent des actions qui peuvent être source d'erreurs. Lorsque l'ordinateur est en cours de traitement, l'utilisateur doit en être informé.

Exemples de recommandations :

- Toujours faire apparaître sur l'écran les entrées effectuées par l'utilisateur, sauf pour les entrées confidentielles.
- Indiquer le curseur actif lorsque plusieurs curseurs sont présents sur une même page écran.

- Toute action à l'initiative de l'utilisateur ou à l'initiative du système doit conduire à un résultat observable.

1.4. Clarté

Définition :

Le critère "Clarté" réfère à tout ce qui concerne les caractéristiques lexicales de présentation des informations sur l'écran (luminance des caractères, contrastes caractères fond, dimensions des lettres, espacements entre les mots, espacements entre les lignes, espacements entre les paragraphes, longueurs des lignes, etc.).

Justification(s) :

La performance est accrue lorsque la présentation des items à l'écran tient compte des caractéristiques cognitives et perceptuelles des utilisateurs. La clarté facilite la lecture des informations présentées. Ainsi par exemple, les lettres sombres sur fond clair sont plus faciles à lire que l'inverse ; le texte présenté en lettres majuscules et minuscules est lu plus rapidement que le texte présenté seulement en lettres majuscules.

Exemples de recommandations :

- Les titres doivent être centrés.
- Les labels doivent être en majuscule.
- Le curseur doit être facilement repérable.

2. Charge de travail

Définition :

La charge de travail concerne l'ensemble des éléments de l'interface qui ont un rôle, pour l'utilisateur, dans la réduction de sa charge perceptive ou mnésique et dans l'augmentation de l'efficacité du dialogue. Deux sous-critères participent à la charge de travail : la "**Brièveté**" qui englobe la "**Concision**" et les "**Actions minimales**", et la "**Charge mentale**".

Justification(s) :

Plus la charge de travail est élevée plus grands sont les risques d'erreurs. De même, moins l'utilisateur sera distrait par des informations non pertinentes, plus il pourra effectuer sa tâche efficacement. Par ailleurs, plus les actions requises seront courtes, plus rapides seront les interactions.

2.1. Brièveté

2.1.1. Concision

Définition :

La concision concerne la charge de travail au niveau perceptif et mnésique en rapport à des éléments individuels d'entrée ou de sortie.

Justification(s) :

Les capacités de la mémoire à court terme sont limitées. Par conséquent, plus courtes sont les entrées, plus petits sont les risques d'erreurs.

Exemples de recommandations :

- Ne pas avoir à entrer les zéros et les blancs devant un item.
- Si des codes sont supérieurs à 4 ou 5 caractères, utiliser des mnémoniques ou abréviations.
- Entrées de données courtes.

2.1.2. Actions minimales

Définition :

Les actions minimales concernent la charge de travail au niveau des options ou moyens utilisés pour atteindre un but.

Justification(s) :

Plus les actions nécessaires à l'atteinte d'un but sont nombreuses et compliquées, plus la charge de travail augmente et par conséquent plus les risques d'erreurs sont élevés.

Exemples de recommandations :

- Minimiser le nombre d'étapes dans la sélection de menus.
- Ne pas entrer de données déjà introduites.
- Eviter les ponctuations pour les entrées de commandes.

2.2. Charge mentale

Définition :

La charge mentale concerne la charge de travail du point de vue perceptif et mnésique, pour des ensembles d'éléments.

Justification(s) :

Dans la plupart des tâches, la performance des utilisateurs est influencée négativement quand la charge informationnelle est trop élevée ou trop faible. La probabilité d'erreurs augmente. Les éléments sans lien avec le contenu de l'écran ne devraient donc pas apparaître, ou du moins apparaître sur d'autres écrans. Il faut donc éviter d'imposer à l'utilisateur la mémorisation de longues et nombreuses informations ou procédures (la mémoire à court terme est limitée), ou toute activité nécessitant de sa part la mise en branle d'activités cognitives complexes.

Exemples de recommandations :

- Limiter la densité de l'information sur l'écran, en affichant que les informations nécessaires.
- L'information ne doit pas nécessiter des traductions d'unités.
- Utiliser le minimum de quantificateurs.

3. Contrôle explicite

Définition :

Le contrôle explicite se réfère à la fois au contrôle qu'a l'utilisateur sur l'interface ou le logiciel, et au caractère explicite de ses actions. Deux sous-critères participent au contrôle explicite : les "**Actions explicites**", et le "**Contrôle utilisateur**".

Justification(s) :

Quand les entrées de l'utilisateur sont explicitement définies par lui-même et sous son contrôle, les ambiguïtés et les erreurs sont limitées. De plus, le contrôle qu'a l'utilisateur sur le dialogue, est un facteur d'acceptation du système.

3.1. Actions explicites

Définition :

Les actions explicites se réfèrent au fait que l'interface doit exécuter seulement les opérations demandées par l'utilisateur.

Justification(s) :

Quand les opérations de l'interface résultent des actions de l'utilisateur, on observe moins d'erreurs.

Exemples de recommandations :

- Pour le remplissage de formulaire, utiliser un "ENTER" explicite à la fin plutôt qu'à chaque entrée.
- Lors d'une sélection d'options de menu par pointage, prévoir une action explicite de sélection.
- L'entrée de commandes doit se terminer par un "ENTER" auxquelles sont préalables des possibilités d'édition.

3.2. Contrôle utilisateur

Définition :

Le contrôle utilisateur se réfère au fait que la plupart de ses actions devraient être anticipées et des options appropriées fournies pour chaque cas. Ceci permet à l'utilisateur d'avoir toujours la main.

Justification(s) :

Quand l'utilisateur a le contrôle de l'interface, les réactions de cette dernière sont prévisibles. L'apprentissage s'en trouve facilité et le risque d'erreurs diminué.

Exemples de recommandations :

- Le curseur ne doit pas être déplacé sans contrôle de l'utilisateur (sauf s'il s'agit de procédures stables et bien connues, e.g. remplissage de formulaires).
- La fin d'une interaction ne doit jamais être déterminée par la position du curseur.
- La vitesse du dialogue doit dépendre de l'utilisateur (ne pas passer d'une page écran à une autre sans contrôle de l'utilisateur).

4. Adaptabilité

Définition :

L'adaptabilité d'un système réfère à sa capacité à réagir selon le contexte, et selon les besoins et préférences de l'utilisateur. Deux sous-critères participent à l'adaptabilité : la **"Flexibilité"** et la **"Prise en compte de l'expérience de l'utilisateur"**.

Justification(s) :

Une interface ne peut convenir à la fois à tous ses utilisateurs potentiels. Pour qu'elle n'ait pas d'effets négatifs sur l'utilisateur, cette interface doit, selon les contextes, s'adapter à l'utilisateur. Par ailleurs, plus les façons d'effectuer une même tâche sont diverses, plus les chances que l'utilisateur puisse choisir et maîtriser l'une d'entre elles, au cours de ses apprentissages, sont importantes. Il faut donc fournir à l'utilisateur des procédures, options, et commandes différentes lui permettant d'atteindre un même objectif.

4.1. Flexibilité

Définition :

La flexibilité concerne les moyens mis à la disposition de l'utilisateur pour personnaliser l'interface afin de rendre compte des exigences de la tâche, de ses stratégies ou habitudes de travail. Elle correspond aussi au nombre de façons différentes mises à la disposition de l'usager pour atteindre un même objectif. Il s'agit en d'autres termes de la capacité de l'interface à s'adapter à des actions variées de l'utilisateur.

Justification(s) :

Plus les façons d'effectuer une même tâche sont diverses, plus les chances que l'utilisateur puisse choisir et maîtriser l'une d'entre elles, au cours de ses apprentissages, sont importantes.

Exemples de recommandations :

- Quand plusieurs entrées de données sont effectuées à la suite, l'utilisateur doit avoir la possibilité de les changer, en tout ou en partie avant de terminer sa séquence d'interaction.
- Quand certains affichages sont inutiles, l'utilisateur doit pouvoir les désactiver temporairement.
- Lorsque des valeurs par défaut ne sont pas connues à l'avance, le système doit permettre à l'utilisateur de sélectionner ces valeurs.

4.2. Prise en compte de l'expérience de l'utilisateur

Définition :

La prise en compte de l'expérience de l'utilisateur concerne les moyens mis en œuvre pour permettre au système de respecter le niveau d'expertise de l'utilisateur.

Justification(s) :

Des utilisateurs expérimentés ont besoin de dialogues moins explicites que les novices. Toutes les commandes ou options n'ont pas à être visibles en tout moment. Des dialogues à la seule initiative de l'ordinateur peuvent ennuyer et ralentir le travail de l'utilisateur expérimenté. Des moyens doivent donc être à la disposition de ce type d'utilisateurs pour leur permettre de contourner ou de s'approprier l'initiative du dialogue.

Exemples de recommandations :

- Prévoir des raccourcis.
- Prévoir des choix d'entrées simples ou multiples selon l'expérience de l'utilisateur.
- Autoriser différents modes de dialogue.

5. Gestion des erreurs

Définition :

La gestion des erreurs concerne tous les moyens permettant d'une part d'éviter ou de réduire les erreurs, et d'autre part de les corriger lorsqu'elles surviennent.
Trois sous-critères participent à la gestion des erreurs : la **"Protection contre les erreurs"**, **"Qualité des messages"**, et la **"Correction des erreurs"**.

Justification(s) :

Les interruptions provoquées par les erreurs ont des conséquences négatives sur l'activité de l'utilisateur. De manière générale, elles rallongent les transactions et perturbent la planification. Plus les erreurs sont limitées, moins il y a d'interruptions au cours de la réalisation d'une tâche et meilleure est la performance.

5.1. Protection contre les erreurs

Définition :

La protection des erreurs concerne les moyens mis en place pour détecter les erreurs d'entrées de données ou de commandes.

Justification(s) :

Il est préférable de détecter les erreurs lors de la saisie plutôt que lors de la validation.

Exemples de recommandations :

- Quand l'utilisateur termine une session et qu'il y a un risque de perte de données, il doit y avoir un message le signalant et demandant confirmation de fin de session.
- Les labels de champs doivent être protégés.
- Les aires d'affichage qui ne sont pas nécessaires pour l'entrée de données ne doivent pas être accessibles à l'utilisateur.

5.2. Qualité des messages

Définition :

La qualité des messages concerne la pertinence et l'exactitude de l'information donnée à l'utilisateur sur la nature de l'erreur commise (syntaxe, format, etc.), et sur les actions à entreprendre pour la corriger.

Justification(s) :

La qualité des messages favorise l'apprentissage du système en indiquant à l'utilisateur la raison ou la nature de son erreur et en lui indiquant ce qu'il faut ou ce qu'il aurait du faire.

Exemples de recommandations :

- Si l'utilisateur sélectionne une touche fonction invalide, aucune action ne doit en résulter, si ce n'est un message indiquant les fonctions appropriées à cette étape de la transaction.
- Fournir des messages d'erreurs orientés tâches.

5.3. Correction des erreurs

Définition :

Concerne les moyens mis à disposition de l'utilisateur pour lui permettre de corriger ses erreurs.

Justification(s) :

Les erreurs sont d'autant moins perturbatrices qu'elles sont faciles à corriger.

Exemples de recommandations :

- Fournir une option retour-arrière.
- Fournir les moyens de résoudre les abréviations ambiguës.
- Fournir une façon d'annuler les effets d'une commande.

6. Homogénéité/Consistence (Consistency)

Définition :

L'homogénéité réfère à la façon avec laquelle des choix d'objets de l'interface (codes, procédures, dénominations, etc.) sont conservés pour des contextes identiques, et des objets différents pour des contextes différents. L'homogénéité s'applique aussi bien à la localisation et au format qu'à la syntaxe et la dénomination.

Justification(s) :

Les procédures, items, sorties, etc., sont d'autant mieux reconnus, localisés et utilisés, que leur format, localisation, ou syntaxe sont stables d'un écran à l'autre, d'une session à l'autre, et d'une application à l'autre. Dans ces conditions le système est davantage prédictible et les apprentissages plus généralisables. Il faut donc adopter des choix similaires de codes, procédures, dénominations pour des contextes identiques, et utiliser les mêmes moyens pour obtenir les mêmes résultats. Il convient de standardiser autant que possible tous les objets quant à leur format et leur dénomination, et standardiser la syntaxe des procédures. Le manque d'homogénéité dans la position des menus par exemple, peut augmenter considérablement le temps de recherche. Le manque d'homogénéité est aussi une raison importante du refus d'utilisation.

Exemples de recommandations :

- Localisation similaire des fenêtres.
- Formats d'écrans similaires.
- Procédures similaires d'accès aux options de menus.

7. Signifiance des codes

Définition :

La signifiance des codes se réfère à l'adéquation entre l'objet ou l'information affichée ou demandée, et son référent.

Justification(s) :

Lorsque le codage est signifiant, le rappel et la reconnaissance sont meilleurs.

Exemples de recommandations :

- Le titre doit véhiculer ce qu'il représente.
- Expliciter la ou les règles de contraction ou d'abréviation.
- Les labels doivent être distincts.

8. Compatibilité

Définition :

La compatibilité réfère à l'accord pouvant exister entre les caractéristiques de l'utilisateur (mémoire, perceptions, habitudes, etc.) et l'organisation des sorties, des entrées et du dialogue.

Justification(s) :

Le transfert d'information est d'autant plus rapide et efficace que le volume d'information à recoder par l'utilisateur est réduit. L'efficacité est accrue lorsque : les procédures nécessaires à l'accomplissement de la tâche sont compatibles aux caractéristiques cognitives de l'utilisateur ; les procédures et les tâches sont organisées de manière à respecter les attentes, ou habitudes de l'utilisateur ; lorsque les traductions, les transpositions, les interprétations, ou références à la documentation sont minimisées. Les performances sont meilleures lorsque l'information est présentée sous une forme directement utilisable (écrans compatibles avec le support papier, dénominations de commandes compatibles avec le vocabulaire de l'utilisateur, etc.).

Exemples de recommandations :

- L'organisation des informations affichées doit être conforme à l'organisation des données à entrer.
- Le format des écrans doit être compatible avec celui des documents papier.
- Les procédures de dialogue doivent être compatibles avec l'ordre tel que se l'imagine l'opérateur ou celui dont il a l'habitude.
- Le format de date en français est jour/mois/année. En anglais, ce format devient : mois/jour/année.

Annexe 2

Matrices de confusions

Tableau 7

Matrice de confusion des critères chez les participant(e)s expérimenté(e)s * **

	Critères identifiés																	
Critères théoriques	1. 1.	1. 2. 1.	1. 2. 2.	1. 3.	1. 4.	2. 1. 1.	2. 1. 2.	2. 2.	3. 1.	3. 2.	4. 1.	4. 2.	5. 1.	5. 2.	5. 3.	6.	7.	8.
1.1.	**18**		*2*[1]		*1*			*1*		*1*			*1*					
1.2.1.		**17**	*2*[2]		*1*						*1*					*3*[3]		
1.2.2.	*10*[8]	*1*	**5**			*1*	*2*[2]			*2*[2]						*1*	*1*	*1*
1.3.	*3*[3]			**20**				*1*										
1.4.		*3*[2]			**20**			*1*										
2.1.1.	*5*[3]					**6**	*9*[6]				*1*							*3*[3]
2.1.2.						*1*	**18**	*1*	*1*		*2*[2]					*1*		
2.2.	*2*[2]					*9*[6]	*1*	**11**									*1*	
3.1.	*1*					*1*	*2*[1]		**16**	*3*[2]								*1*
3.2.	*1*					*1*	*1*		*4*[4]	**11**	*4*[3]		*1*		*1*			
4.1.							*2*[2]				**18**							*4*[4]
4.2.							*5*[4]				*8*[7]	**11**						
5.1.	*4*[3]									*1*			**18**				*1*	
5.2.														**23**			*1*	
5.3.	*2*[2]									*6*[4]	*3*[3]		*1*		**11**		*1*	
6.	*1*	*1*				*1*							*1*			**20**		
7.	*2*[2]	*1*							*1*						*1*	*2*[2]	**15**	*2*[2]
8.	*5*[5]	*1*									*1*					*1*		**16**
Total	54	24	9	20	22	20	40	15	22	24	38	11	22	23	13	28	20	27
Corr.[1]	18	17	5	20	20	6	18	11	16	11	18	11	18	23	11	20	15	16
Incorr.[2]	36	7	4	0	2	14	22	4	6	13	20	0	4	0	2	8	5	11
Théorique[3]	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24

* Les nombres apparaissant en caractères gras sur la diagonale constituent le nombre d'accords.

** Les nombres apparaissant en caractères italiques constituent des désaccords. Les exposants qui accompagnent les fréquences supérieures à 1 représentent le nombre de participant(e)s qui y concourrent.

1 Nombre d'identifications correctes.

2 Nombre d'identifications incorrectes.

3 Nombre d'identifications théoriquement correctes.

Tableau 8

Matrice de confusion des critères chez les participant(e)s inexpérimenté(e)s * **

Critères théoriques	Critères identifiés																	
	1.1.	1.2.1.	1.2.2.	1.3.	1.4.	2.1.1.	2.1.2.	2.2.	3.1.	3.2.	4.1.	4.2.	5.1.	5.2.	5.3.	6.	7.	8.
1.1.	**13**	*1*	*3*^{3}		*2*^{2}	*1*				*1*		*1*					*1*	*1*
1.2.1.	*1*	**14**	*5*^{4}					*1*				*1*				*1*	*1*	
1.2.2.	*4*^{3}	*1*	7		*1*	*1*	*1*		*1*	*1*		*1*				*4*^{4}	*2*^{1}	
1.3.	*6*^{5}			**14**		*1*		*2*^{2}			*1*							
1.4.			*1*		**20**			*2*^{1}								*1*		
2.1.1.	*3*^{3}					7	*8*^{6}					*1*					*4*^{3}	*1*
2.1.2.	*1*						**11**		*1*	*5*^{4}	*6*^{5}							
2.2.	*1*		*1*		*5*^{5}	*5*^{4}	*1*	**11**										
3.1.						*2*^{2}			**15**	*6*^{6}			*1*					
3.2.				*3*^{3}			*1*	*1*	*1*	**14**			*1*		*3*^{2}			
4.1.	*1*								*1*	*1*	**17**	*1*			*2*^{2}			*1*
4.2.							*1*		*1*	*2*^{1}	*11*^{8}	**8**					*1*	
5.1.										*1*		*1*	**18**		*2*^{2}	*1*	*1*	
5.2.						*1*			*1*				*1*	**19**			*2*^{2}	
5.3.	*1*									*9*^{7}	*1*		*1*		**12**			
6.		*1*					*1*	*1*	*1*							**20**		
7.	*2*^{2}	*5*^{5}	*1*			*1*			*1*				*1*		*1*		**12**	
8.		*1*									*2*^{2}					*2*^{2}	*3*^{2}	**16**
Total	33	23	18	17	28	19	24	18	23	40	38	14	23	19	20	29	27	19
Corr.[1]	13	14	7	14	20	7	11	11	15	14	17	8	18	19	12	20	12	16
Incorr.[2]	20	9	11	3	8	12	13	7	8	26	21	6	5	0	8	9	15	3
Théorique[3]	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24

* Les nombres apparaissant en caractères gras sur la diagonale constituent le nombre d'accords.

** Les nombres apparaissant en caractères italiques constituent des désaccords. Les exposants qui accompagnent les fréquences supérieures à 1 représentent le nombre de participant(e)s qui y concourrent.

1 Nombre d'identifications correctes.

2 Nombre d'identifications incorrectes.

3 Nombre d'identifications théoriquement correctes.

Imprimé en France
par
. l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249 - 6399